Aalto University
School of Electrical
Engineering

AALTO UNIVERSITY
School of Electrical Engineering
Department of Communications and Networking

Rabiul Islam Jony

Preprocessing Solutions for Telecommunication Specific Big Data Use Cases

Master's Thesis submitted in partial fulfillment of the degree of Master of Science in Technology

Espoo, November, 2013

Supervisor: Prof. Heikki Hämmäinen, Aalto University, Finland

Instructor: Lic.Sc. (Tech.) Pekka Kekolahti, Aalto University, Finland

AALTO UNIVERSITY **ABSTRACT OF THE MASTER'S THESIS**

Author:	Rabiul Islam Jony
Title of the Thesis:	Preprocessing Solutions for Telecommunications Specific Big Data Use cases
Date:	28.11.2013 Number of pages: 12+65
School:	School of Electrical Engineering
Department:	Department of Communications and Networking
Professorship:	Network Economics
Supervisor:	Prof. Heikki Hämmäinen, Aalto University, Finland
Instructor:	Lic.Sc. (Tech.) Pekka Kekolahti, Aalto University, Finland

Big data is becoming important in mobile data analytics. The increase of networked devices and applications means that more data is being collected than ever before. All this has led to an explosion of data which is providing new opportunities to business and science. Data analysis can be divided in two steps, namely preprocessing and actual processing. Successful analysis requires advanced preprocessing capabilities. Functional needs for preprocessing include support of many data types and integration to many systems, fit for both off-line and on-line data analysis, filtering out unnecessary information, handling missing data, anonymization, and merging multiple data sets together.

As a part of the thesis, 20 experts were interviewed to shed understanding on big data, its use cases, data preprocessing, feature requirements and available tools. This thesis investigates on what is big data, and how the organizations, especially telecommunications industry can gain benefit out of it. Furthermore, preprocessing as a part of value chain is presented and the preprocessing requirements are sorted. Finally, The available data analysis tools are surveyed and tested to find out the most suitable preprocessing solution.

This study presents two findings as results. Firstly, it identifies the potential big data use cases and corresponding functional requirements for telecom industry based on literature review and conducted interviews. Secondly, this study distinguishes two most promising tools for big data preprocessing based on the functional requirements, preliminary testing and hands-on testing.

Keywords: *Big data, Data preprocesing, Data value chain, Big data in telecom industry, Preprocessig tools.*

Language: English

ACKNOWLEDGMENTS

This master's thesis has been carried out in the Department of Communications and Networking of Aalto University School of Electrical Engineering as a part of the research project 'MOMIE – Modeling of Mobile Internet Ecosystem'.

First of all, I would like to express my sincere gratitude to my supervisor, Professor Heikki Hämmäinen for giving me the opportunity to work under his supervision and guiding me throughout the process.

My heartfelt thank and appreciation goes to my instructor Lic.Sc. (Tech.) Pekka Kekolahti, for his guidance, encouragements, ideas, advice and endless patient on me. Its been a pleasure working with him.

My special thanks to all the experts who participated in the interview sessions, for their kind cooperation and valuable time.

Strongest thanks to my parents for their continuous support, encouragement and blessing at every step of my life.

I am grateful to Rashedul Islam Khan Rana, Rakibul Islam Rony and Taslima Akter Shoma for their enormous support and courage on my working and being aside of me always.

Special thanks to Youtube, and the online forums of the tools.

Otaniemi, Espoo: 28.11.2013

Rabiul Islam Jony

To my Parents

Table of Contents

ABBREVIATIONS

AMA	Automatic Message Accounting
ARFF	Attribute-Relation File Format
AVG	Average
BSC	Base Station Controller
BSS	Business Support Systems
BTS	Base Transceiver Station
CAGR	Compound Annual Growth Rate
CDR	Call Detail Record
CRM	Customer Realtionship Management
CSV	Comma-separated values
DBA	Database Administator
DDDM	Data-Driven Decision-Making
ELT	Extract-Load-Transfrom
ETL	Extract-Transfor-Load
GPS	Global Positioning System
GUI	Graphical User Interface
HDFS	Hadoop Distributed File System
HLR	Home Location Register
HTTP	Hypertext Transfer Protocol
IAP	Internet Access Provider
IoT	Internet of Things
IP	Internet Protocol
IPDR	Internet Protocol Detail Record
IPG	Inter Packet Gap
ITU	International Telecommunication Union
LAC	Location Area Code
LTE	Long-Term Evolution
M2M	Machine-to-Machine
MAC	Media Access Control
Max	Maximum
MCC	Mobile Country Code
Min	Minimum
MMS	Multimedia Messaging Service
MNC	Mobile Network Code
MSC	Mobile Switching Center
NA	Not Available/Not Applicable
NoSQL	Not only Structured Query Language
NR	Not Ranked
OSS	Operation Support Systems
OTT	Over-The-Top
PCA	Principal Component Analysis
PdM	Predictive Maintenance
QoE	Quality of Experience
QoS	Quality of Service
RAID	Redundant Array of Independent Disks
RBS	Radio Base Station

RFC	Request for Comments
RNC	Radio Network Controller
RTCP	Real-time Transport Control Protocol
RTP	Real-time Transport Protocol
SDR	Service Detail Record
SIP	Session Initiation Protocol
SLA	Service-level Agreement
SMS	Short Message Service
SON	Self-Organizing Network
SQL	Structured Query Language
SVM	Support Vector Machine
TCP	Transmission Control Protocol
TXT	Text files
UDP	User Datagram Protocol
VLR	Visitor Location Register
VOD	Video on Demand
XDR	Extended Data Record
XML	Extensible Markup Language
XSD	XML Schema

List of Figures

List of Tables

1. Introduction

1.1. Background

In 2000, when the Sloan Digital Sky Survey started their operation, its telescope in New Mexico collected more data on its first few weeks than had been amassed in the entire history of astronomy. After one decade, its archive presently contains around 140 terabytes of data. Another large Synoptic Survey Telescope in Chile is predicted to collect a corresponding quantity of data every five days by 2016 (The Economist, 2010) . The retail giant Wal-Mart's warehouse stores around 2.5 petabytes of data regarding 1 million customer transactions on an hourly basis (Infosys, 2013). Facebook, a social networking website stores 500+ terabytes of new data every day. Search engines, such as Google daily process 20 petabytes of data (The Economist, 2010). All these examples show the big amount of data the world contains, and the speed at which the volume of data is growing. The mankind had managed to create 5 exabytes of data by 2003, and today the same amount of data is created in only two days (Schmidt, 2010). The amount of data in the digital world reached 2.72 zettabytes in 2012, and is expected to double every two years reaching 8 zettabytes by 2015 (Sagiroglu & Sinanc, 2013). Data are getting so large and complex, that it is becoming difficult to process using traditional data processing applications, and introducing big data.

The most popular definition of big data is defined by Gartner as "Big data is high-volume, high-velocity and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" (Beyer & Laney, 2012). Definitions of big data will be discussed in more detail in Chapter 2 of this thesis.

The telecommunications industry constantly transfers petabytes of data across their networks. Due to next generation mobile network rollouts, such as LTE (Long-term evolution), IoT (internet of things), and M2M (Machine-to-Machine communications) along with increased use of smart phones and rise of social media, mobile operators are experiencing a rise in volume, variety and velocity of data. According to (CITO Research, 2012), 54 percent of operators claimed, that big data is a current strategic priority in their organization. According to ITU, 2013 will be the year, when the majority of Mobile operators will start treating big data as strategic priority for both internal and external use (ITU, 2013).

Big data is currently treated as a technology, which has been developed to handle large volumes of fast-changing and non-schematic data. Big data technology also provides companies, such as telecom operators with an ideal platform for centralizing and storing and analyzing their structured,

unstructured and semi-structured data. These yield major advantages in data analysis, knowledge discovery and new business opportunity identification.

Data analysis is divided into two steps, namely preprocessing and actual processing. Successful processing requires advanced preprocessing capabilities. The data which are not yet subjected to any analysis are known as raw data. Data preprocessing can be defined as bringing out the right data from the raw data, or preparing the raw data for actual processing.

Many tools are presently available for data analysis, new tools focusing on big data analytics are also emerging. Due to new internet business models, different open source preprocessing and analytics tools have been developed which might be well suited for operators' data preprocessing.

1.2. Motivation

Operators are collecting large amounts of data every day. The collected data can provide valuable information for the operators about subscriber experience, e.g. how the call has gone through, whether it was dropped or interrupted, how fast the apps were downloaded, and how was the response latency. Collected data also allow the operators to learn about the interest of a subscriber, e.g. which websites the subscriber visits most, what kind of application is being downloaded and what is the subscriber sentiment conveyed in the social media. Network data and other external data can also provide valuable information to the operators which can be applied in different use cases.

This thesis has been made in Network Economics Research Group of Prof. Heikki Hämmäinen in the School of Electrical Engineering, Aalto University. The research team has vast experience in data analysis as part of the mobile ecosystem research. The tools and libraries the researchers of this research group are using include Excel, R, Matlab, Weka, SPSS, and Bayesialab. The datasets they are working on include handset based monitoring data, survey data, device databases, sales data, traffic data (TCP/IP, HTTP). Although the mentioned tools and libraries have some level of preprocessing capabilities, new emerged open source tools with modern user interfaces and rich functionalities would facilitate the group's research tasks and extend the scope and possibilities as increasing data sources will be available for them.

1.3. Research Questions

This thesis attempts to answer two questions relevant to the big data in telecom industry and Preprocessing of data including:

- Q1: What are the potential big data use cases in telecom industry?

- Q2: Which of the available tools support best the functionality and usability, regarding telecom industry data preprocessing tasks (including small and big data preprocessing)?

1.4. Problem statement

Big data is enabling new business cases for the mobile operators, and it is important for them to find out the potential use cases. Operators for example also need to know what are the typical data types, data sources and requirements for the use cases. The data preprocessing is a vital part of the data value chain. It is important in finding out, when data needs to be reduced, cleaned and modified. The preprocessing feature requirements and capable tools-list are important while choosing proper tools out of many. Open source tools are emerging, and it is important to know how these tools can perform data preprocessing, also in academic work.

1.5. Objective of the thesis

This thesis will sort out potential big data use cases for operators. The data types, data sources, and the requirements for those specific use cases will also be presented. The feature requirements for the preprocessing of data will be classified and selected tools will be graded according to their capability. Finally, several preprocessing tasks with typical datasets will be accomplished with illustrious tools to find out the most suitable tools.

The objectives of this thesis are:

- To discover potential big data use cases for telecom operators, and to list the functional requirements for the tools based on the use cases.
- To figure out the typical challenges in preprocessing of big data
- To understand how the tools meet the preprocessing requirements
- To find out how the most promising tools can perform in practical preprocessing tasks, through practical hands-on testing.

As a part of the thesis, 20 experts form Aalto University, vendor, operator and media companies were interviewed to shed understanding on big data, its use cases, data preprocessing, feature requirements and available tools.

1.6. Outline of the thesis

In Chapter 2 of this thesis, big data will be described and defined, and big data value chain will be discussed.

Chapter 3 deals with the typical data types and data sources for telecom operators. The potential big data use case domains will be shortly described along with few examples.

Chapter 4 will be focused on the data preprocessing as a process. Major tasks and techniques of preprocessing will be discussed.

In Chapter 5, the available data preprocessing tools will be listed. The feature requirements for data preprocessing will also be described.

In Chapter 6, the test cases will be introduced for the hands-on testing of top scored tools and the results will be discussed.

Chapter 7 concludes the thesis with discussing the assessment, exploitation, and future research opportunities of this study.

2. Big data

2.1.*Definition*

Big data has been defined simply as "Big data refers to data volumes in range of exabytes (10^{18}) and beyond" in (Kaisler, et al., 2013).

According to Wikipedia, "Big data is a collection of datasets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications, where the challenges include capture, storage, search, sharing, transfer, analysis, and visualization" (Wikipedia, 2013). In this definition big data is addressed as a problem.

Sam Madden from Massachusetts Institute of Technology (MIT) wrote "Big data means too big, too fast, or too hard for existing tools to process" (Madden, 2012). He also explained, the term 'too big' as the amount of data which might be at petabyte-scale and come from various sources, 'too fast' as the data growth, which is fast and must be processed quickly, and 'too hard' as the difficulties of big data that does not fit neatly into an existing processing tool (Madden, 2012).

From PC Mag (popular magazine based on latest technology news), "Big data refers to the massive amounts of data that collects over time that are difficult to analyze and handle using common database management tools" (PC Magazine, 2013).

John Weathington has defined big data as a competitive key parameter in different dimensions such as customers, suppliers, new entrants and substitutes. According to him, big data creates products which are valuable and unique, and prelude other products from satisfying the same need. He also described, "Big data is traditionally characterized as a rushing river: large amounts of data flowing at a rapid pace" (Weathington, 2012).

It has also been defined as "Big data Represents the progress of the human cognitive processes, usually includes data sets the sizes beyond the ability of current technology, method and theory to capture, manage, and process the data within a tolerable elapsed time" (Doctorow, 2008).

Philip Hunter in has stated, "Big data embodies an ambition to extract value from data, particularly for sales, marketing, and customer relations" (Hunter, 2013).

Svetlana Sicular has defined big data as "high-volume, -velocity and –variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" (Sicular, 2013).

There are many more big data definitions available describing the different characteristics of it.

2.1.1. Big data characteristics

The characteristics of big data are well defined in the definition by Gartner (Beyer & Laney, 2012). The three Vs (volume, velocity and variety) are known as the main characteristics of big data.

The characteristics are described below.

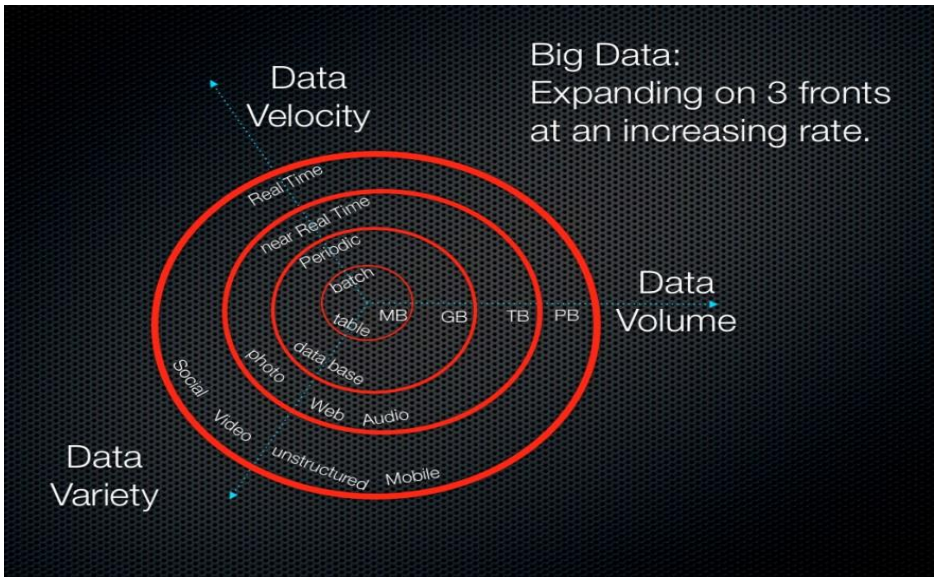

Figure 1: 3Vs of big data **(**Soubra**,** 2012**)**

- **Volume**: Data volume measures the amount of data available to an organization; the organization does not necessarily have to own all of it as long as it can access it (Kaisler, et al., 2013). The number of sources of data for an organization is growing. More data sources consisting large datasets increase the volume of data, which needs to be analyzed. As data volume increases, the value of different data records decreases in portion to age, type, richness and quality among the other factors (Kaisler, et al., 2013).

  Figure 1 shows that the data volume is growing from megabytes (10^6) to petabytes (10^{15}) and beyond.

  Figure 2 indicates, that the volume of data stored in the world would be more than 40 zettabytes (10^{21}) by 2020 (AT Kearney, 2013).

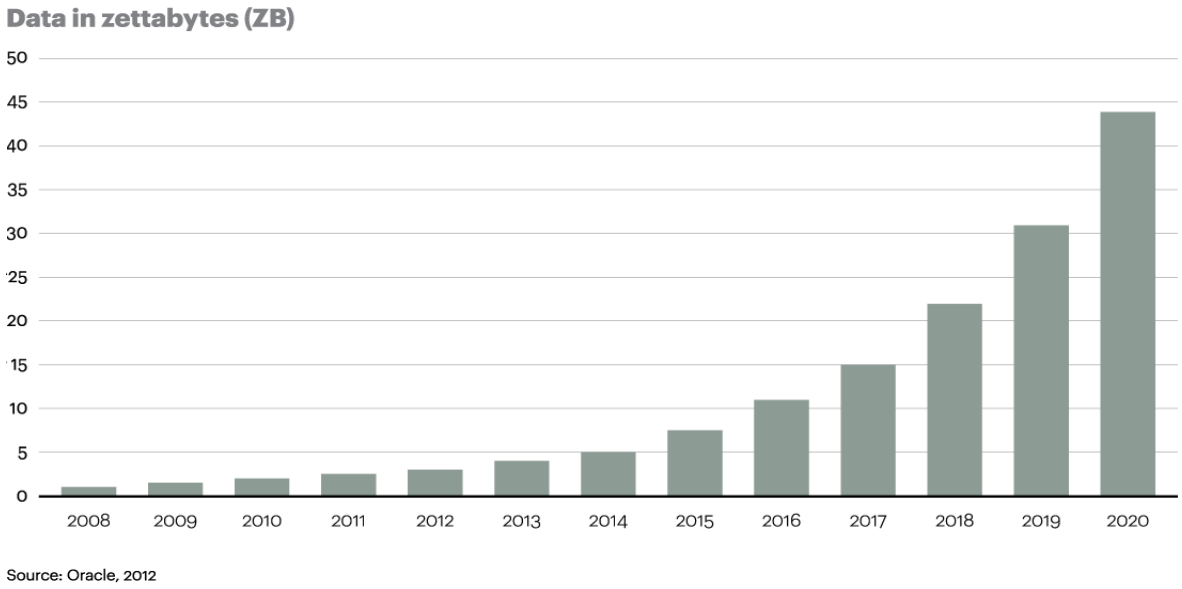

Figure 2: Data volume growth by year in zettabytes (AT Kearney, 2013)

- **Velocity:** Data velocity measures the speed of data creation, streaming and aggregation (Kaisler, et al., 2013). According to Svetlana Sicular from Gartner, velocity is the most misunderstood big data characteristic (Sicular, 2013). She describes that the data velocity is also about the rate changes, and about combining data sets that are coming with different speeds. The velocity of data also describes bursts of activities, rather than the usual steady tempo where velocity frequently equated to only real-time analytics (Sicular, 2013).

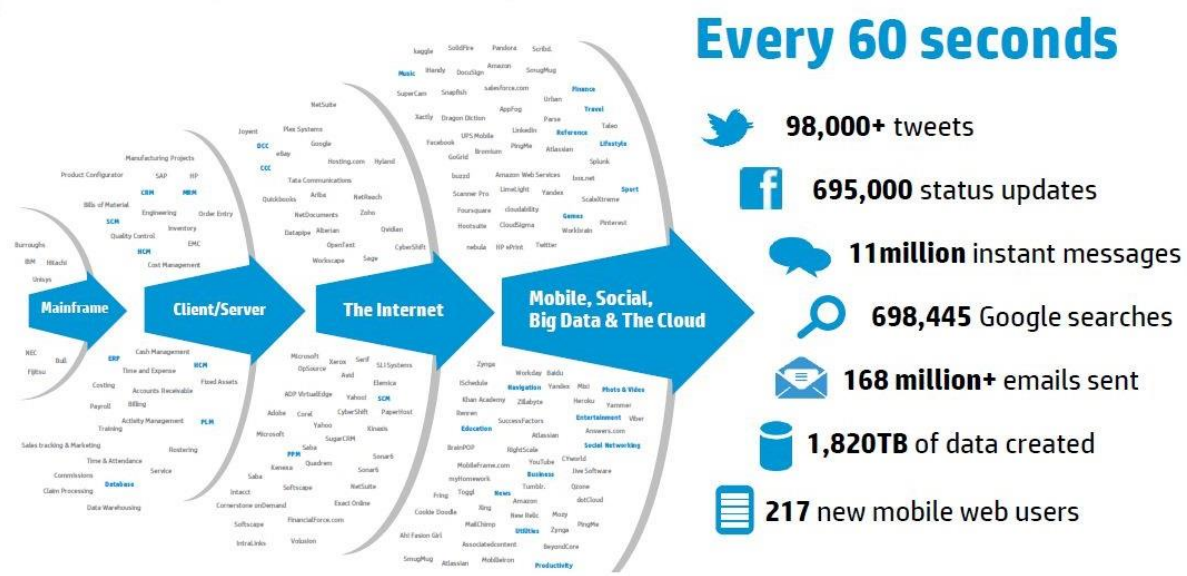

Figure 3: Examples of big data velocity **(**Kalakota**,** 2012**)**

Figure 3 shows few examples of the pace the data. Data velocity management is much more than a bandwidth issue; it is also an ingest issue (Kaisler, et al., 2013).

Figure 1 also reflects velocity as a characteristic of big data, showing how it requires near real-time and/or real-time analytics.

- **Variety:** Other than typical structured data, big data contains text, audio, images, videos, and many more unstructured and semi-structured data, which are available in many analog and digital formats. From an analytics perspective, variety of data is the biggest challenge to effectively use it. Some researchers believe that, taming the data variety and volatility is the key of big data analytics (Infosys, 2013). Data variety is a measure of the richness of the data presentation. Incomputable data formats, non-aligned data structures and inconsistent data semantics represents significant challenges that can lead to analytic sprawl (Kaisler, et al., 2013).

  Figure 4 shows the comparison between increment of unstructured, semi-structured data and structured data by years. Figure 1 also reflects the increment in verity of data.

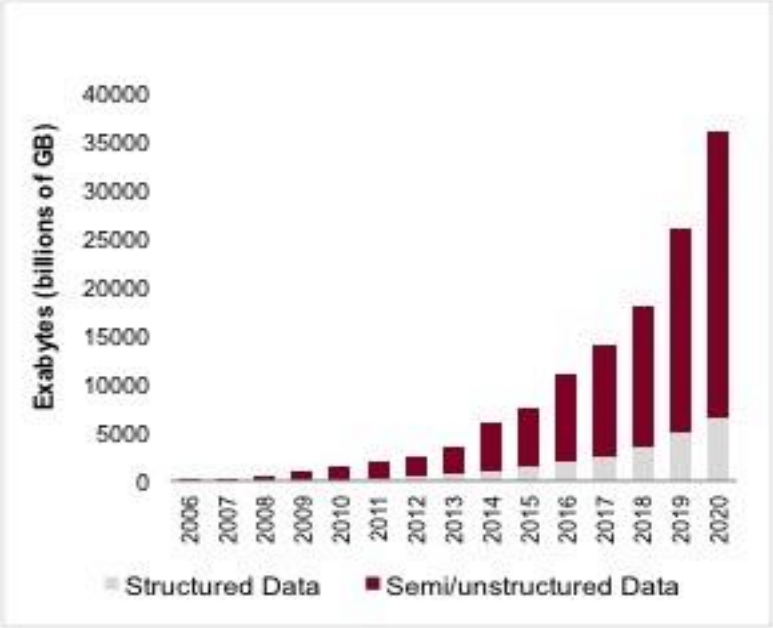

Figure 4: Growth of data variety by years **(**Botteri**,** 2012**)**

One of the big data vendors, IBM has coined additional V for the big data characteristics, which is veracity. By veracity, they address the inherent trustworthiness of the data. As big data will be used e.g. for decision making, it is important to make sure that the data can be trusted.

Some researchers mentioned ‘viability’ and ‘value’ as the fourth and the fifth characteristics leaving ‘veracity’ out (Biehn, 2013).

Figure 5: Four characteristics (volume, velocity, variety, and veracity) of big data **(**IBM**,** 2012**)**

The characteristics of big data can also be described with HACE theorem. The theorem states that, "Big data starts with large-volume; **h**eterogeneous, **a**utonomous sources with distributed and decentralized control and seeks to explore **c**omplex and **e**volving relationships among data (Wu, et al., 2013). From the theorem the key characteristics are defined as:

2. **Huge Data with Heterogeneous and Diverse Dimensionality**: Here the 'heterogeneous' feature refers to the different types of representations for the same individuals. The feature 'diverse' reflects the variety of the features involved to represent each single observation.
3. **Autonomous Sources with Distributed and Decentralized Control**: 'Autonomous' feature describes the ability of each data sources to generate and collect information without any centralized control.
4. **Complex and Evolving Relationships**: With volume of data the complexity and the correlations among them increases.

In summary, big data can be defined as large volume, high velocity and verities of data, which is complex to process with traditional applications, but able to bring new business opportunities to the industries by enhanced insight generation.

Table 1 below summarizes the definitions for big data.

Definitions	Challenging for traditional applications/ Requires new forms of application	Large volume of data	Competitive key parameter	Enhanced insights generator	High-volume, high-velocity, high-variety	Volume, velocity, variety, veracity	HACE Theorem
(Kaisler, et al., 2013)		x					
(Wikipedia, 2013)	x						
(Madden, 2012)	x				x		
(PC Magazine, 2013)	x						
(Weathington, 2012)			x		x		
(Doctorow, 2008)	x						
(Hunter, 2013)				x			
(Sicular, 2013)				x	x		
(IBM, 2012)						x	
(Wu, et al., 2013)							x

Table 1: Definitions for big data

2.2.*Big data business opportunities in different industries*

In 1974, economist John Kenneth Galbraith stated, that performing complex tasks require a greater amount of data to be processed. He also mentioned 'vertical information systems' as the technologies that enable greater collections of information/data (Brynjolfsson, et al., 2011). These facilitate more efficient distribution of information within an organization which lessen the costs and improve the performance.

According to McKinsey Global Institute (MGI) research, big data is becoming the key basis of competition, underpinning new waves of productivity growth, innovation and customer surplus of the future market (McKinsey & Company, 2011).

Figure 6 presents the amounts of stored data in petabytes by industry types in the United States in 2009 (McKinsey & Company, 2011). It shows that industries, such as Manufacturing, Government, Communications and media, and Banking have more the 500 petabytes of data already stored in their systems. Healthcare, Security, Professional services and Retail industries are holding more than 350 petabytes of data. Industries like Education, Insurance, Transportation, Wholesale and Utilities have more than 200 petabytes of data.

Telecommunications industry, combined with other communications and media industry holds the third position in the ranking having around 715 petabytes of stored data.

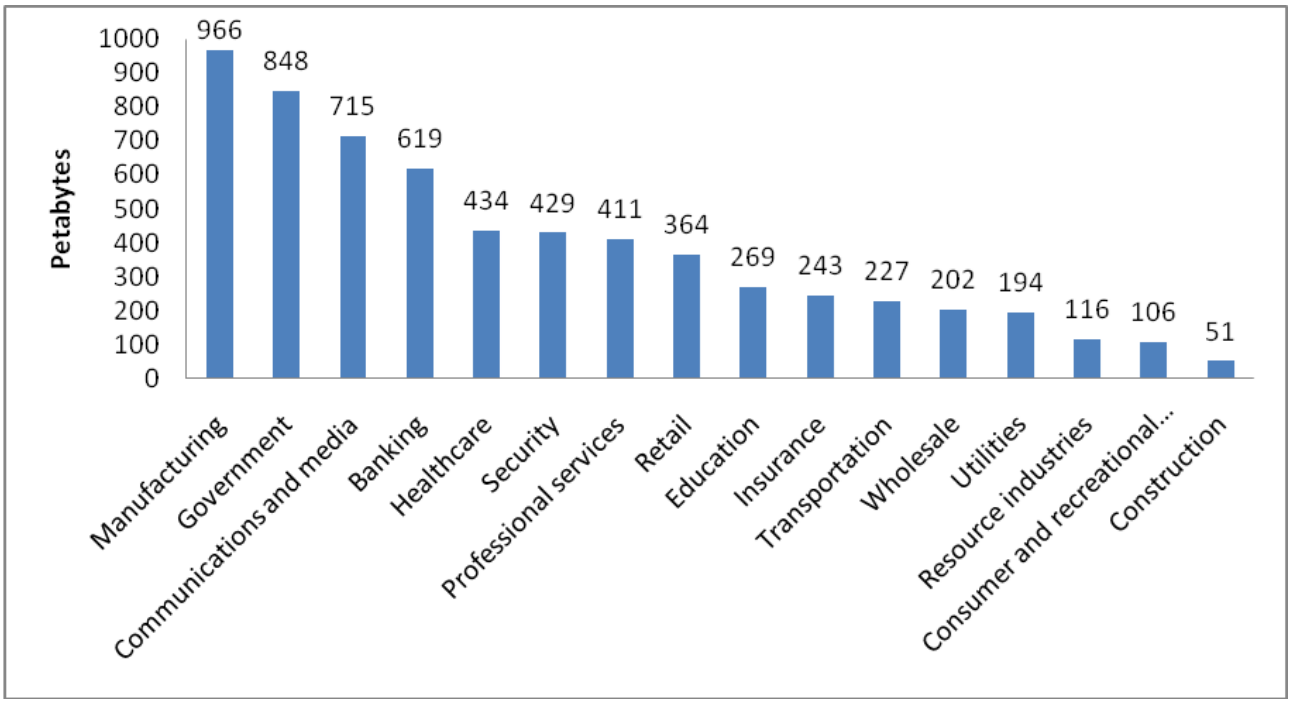

Figure 6: Amount of stored data by industry types in the United States, 2009

The cloud architecture, open source software and commodity hardware of recent market made the big data processing available to the companies which are not even highly resourced. Researchers argued that big data is not just a property of big web companies like Google or Facebook, organizations of all different sizes and industry groups are now leveraging it in many ways (McKendrick, 2013).

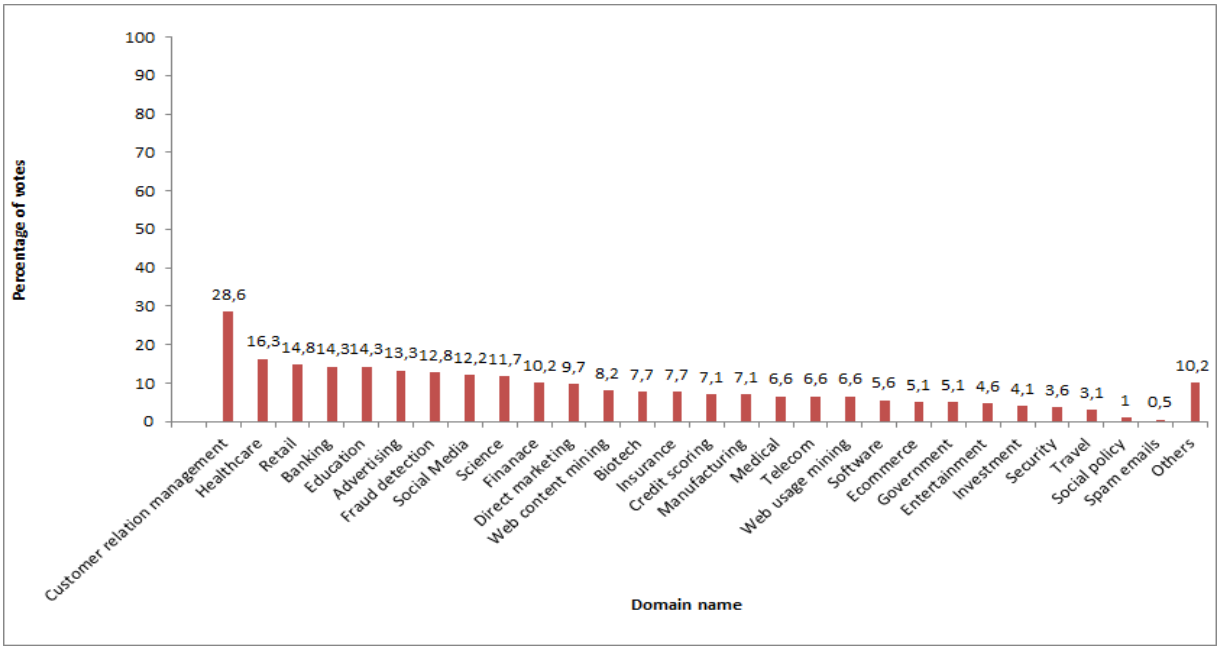

Figure 7: Big data potentiality in different industry

Figure 7 shows the potentiality of big data as a new business asset in different industries (KDnuggets, 2012). It shows that the top industries or domain types are customer relation management and healthcare, followed by retail, banking, education, advertising, fraud detection and so on. Telecom industry is in the latter half in the ranking. Telecom industry, in practice includes other domains, such as customer relationship management, fraud detection and social media. This fact reflects the potentiality of big data in telecom industry as well.

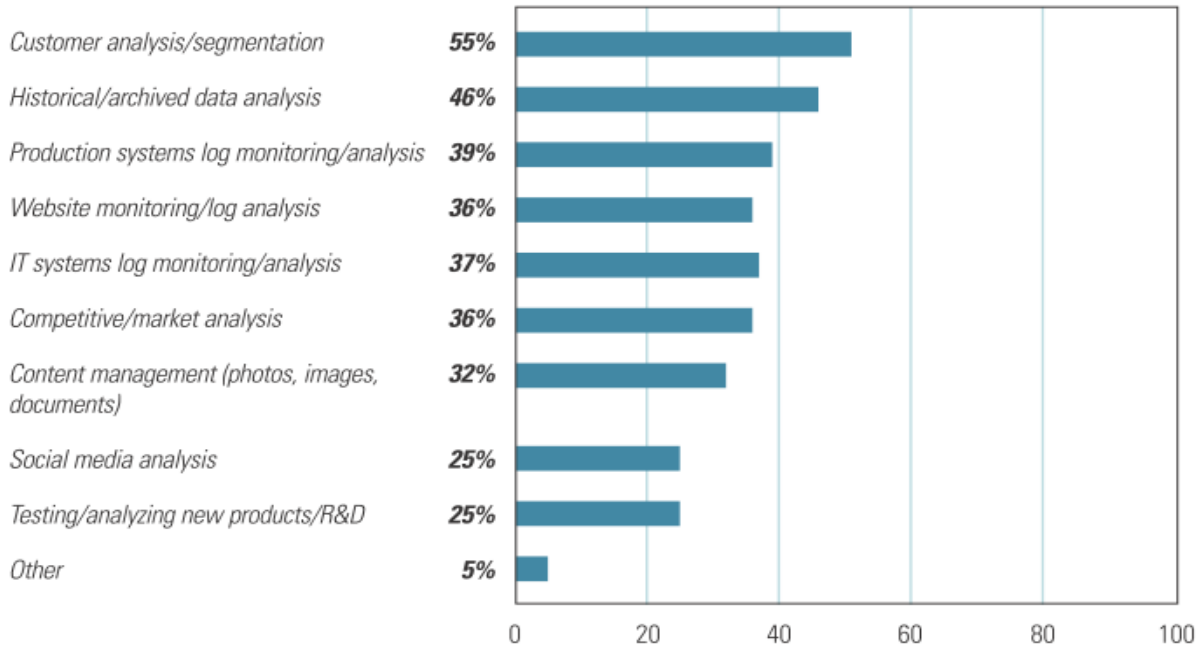

Figure 8: Types of big data initiatives within an organization **(**McKendrick**,** 2013**)**

Figure 8 lists big data initiatives emerged from the survey explained in (McKendrick, 2013). According to the survey, the two top most initiatives are customer analysis or segmentation and historical data analysis. Other initiatives includes production system log monitoring, IT system log monitoring, market analysis and so on.

Big data can bring business benefits to the organizations in several ways; few applicable ways are discussed below.

- **Create transparency**: Organizations can make their big data accessible for their stakeholders in a timely way and make their business progress and strategy transparent.
- **Supply chain management**: Big data analytics can improve organizations' supply chain management by ensuring real-time delivery management, better vendor management, automated product sourcing and segmented supply chain.

- **Allow experimentation to discover needs**: Big data analytics will allow the organizations to do more experiments on their business strategy and products to find out new requirements.
- **Dynamic pricing**: Big data analytics allow the organizations to optimize their product pricing according to the customer needs, market research and business target.
- **Targeted marketing**: Big data can provide the organizations with information about customers' needs and interests, which will allow them to do targeted marketing.
- **Innovate new business models, services and products**: Proper data analysis can provide the organizations with insights about the market, e.g. which types of products are being appreciated or neglected by the customers. New business models, new services or products can be invented by analyzing these types of data.
- **Predictive analysis**: Predictive analysis can identify events before occurring and predict the outcome before implementation.
- **Better understand the customers**: Organizations can utilize their customer data to understand their customers more efficiently and know how, when, and what they want.
- **Lessen Business OPEX and CAPEX**: Big data analytics will allow the organizations to lessen their operational cost and capital cost in different ways, e.g. by optimizing business strategy.
- **Improve performance**: Big data analysis can also offer improved performance by improving decision making accuracy and saving time.
- **Compete more effectively**: Big data analysis will allow the organizations to know about themselves and the rivals in the market. Proper Big data analysis will allow the companies to compete more effectively and survive in the market.
- **Data-Driven Decision-Making (DDDM)**: DDDM is the technology to use data analytics as insights for decision making. Survey on 179 large publicly traded firms showed, the firms that adopt DDDM had output and productivity as high as 6% than the others who did not (Brynjolfsson, et al., 2011). More data analysis will offer more effective and accurate decision making.
- **Insight generation**: Analyzing big data will allow the organizations to generate more insights, which were not possible with small datasets.

Big data also has some limitations. Sometimes it can produce few patterns which are entirely caused by chance, not replicable or having no predictive power. It might also provide weaker patterns, where strong patterns get ignored. In (Boyd & Crawford, 2011), six provocations for big data have

been discussed. The study argued, that the term ‘Big data’ is not perfect for it, as it is notable not because of its size, but because of its confliction with other data. According to the study, big data is still subjective, clamming objectivity and accuracy of it is misleading. The study has also argued that bigger data are not always better data; meaning that the quantity does not necessarily mean quality and it is important to focus and learn the value of small data first. The study describes one provocation of big data as limited access to big data creates new digital divides. Finally, big data analytics also have potential privacy and ethical issues, because it does not make it ethical just because it is accessible (Boyd & Crawford, 2011).

A proper big data value chain can facilitate the organizations to get the business benefits in above described ways.

2.3.*Big data value chain*

Few decades ago, Michale E. Porter first introduced the concept of value chain, where he explained a value chain as a series of activities that create and build value as it progresses (Porter, 1985). Finally these activities culminated in total value, which the organizations then deliver to its customer (Miller & Mork, 2013). In 1988 R. L. Ackoff first specified data value chain (Ackloff, 1989). This was a hierarchy based on filtration, reduction, and transformation showing how data lead to information, knowledge and finally to wisdom. He presented Data-Information-Knowledge-Wisdom hierarchy as a pyramid which produces a series of opposing terms including misinformation, error, ignorance and stupidity when inverted (Bernstein, 2011). Ackoff has fitted wisdom on the top of the hierarchy pyramid followed by knowledge, information and then the data or the raw data.

Figure 9: The Data-Information-Knowledge-Wisdom hierarchy pyramid

Table 2 below describes the four components of Data-Information-Knowledge-Wisdom hierarchy.

Category	Description
Data	Data is raw. It simply exists and has no significance beyond its existence and it does not have any meaning of itself (Bellinger, et al., 2004). Data can also be defined as Computerized representation of models and attributes of real or simulated entities (Chen, et al., 2008).
Information	Information is the data that has been given meaning by way of relational connection (Bellinger, et al., 2004). Information can also be defined as the data that represents the results of the computational process such as statistical analysis, providing answers to questions, such as 'who', 'what', 'where' and 'when'.
Knowledge	Knowledge is the appropriate collection of the information calculated out of raw data and its intent has to be useful. Knowledge might also be defined as the data that represents the results of a computer-simulated cognitive process, such as perception, learning, and reasoning. Knowledge is the application of data and information which provides the answers to 'how' questions (Chen, et al., 2008).
Wisdom	Wisdom represents the ability to see the long-term consequences of any act and evaluate them relatively to the ideal of total control. Wisdom is a non-deterministic and non-probabilistic process that answers questions like 'what needs to be done and why'. Wisdom can also be defined as the process by which the outcome can be judged (Ackloff, 1989).

Table 2: Data-Information-Knowledge-Wisdom components

The big data value chain in this research is divided into three steps, naming Data sources, Preprocessing and storing, and Processing and Visualization, where each step increases value.

2.3.1. Data sources, types and accessibility

The data types and accessibility are included in the sources tag because these also define the value. This step can be divided into three sub-divisions naming availability, amount and accessibility. These define the value of the data sources.

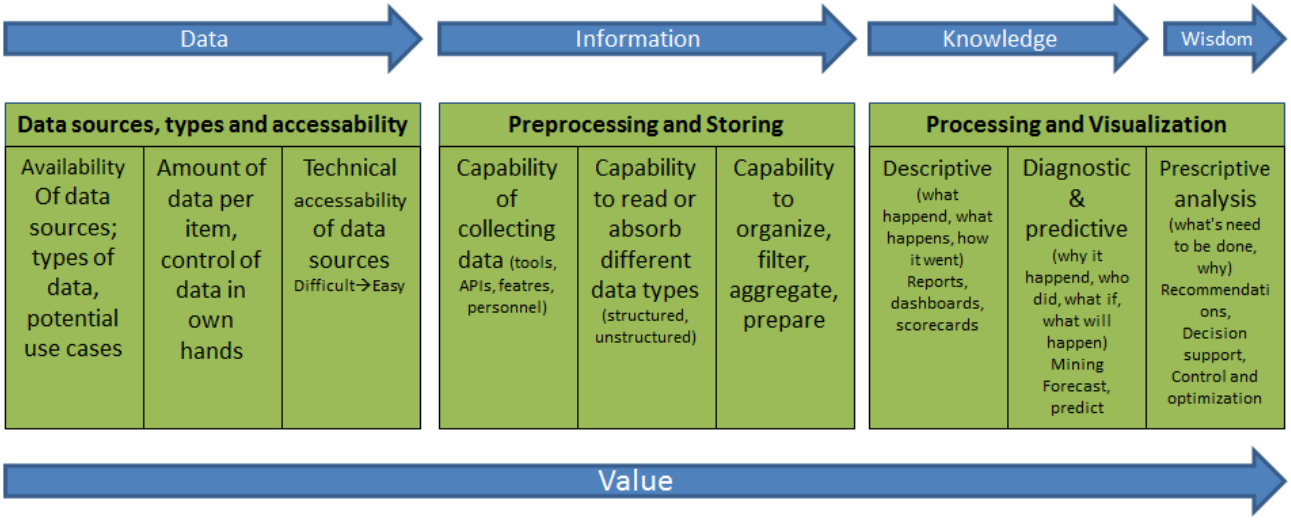

Figure 10: Typical big data value chain

In Figure 10, difficult→easy is mentioned in this step, which means if the data from the sources is easily accessible, it has higher value. This step of the value chain lies under the data section of the Data-Information-Knowledge-Wisdom pyramid.

2.3.2. Preprocessing and Storing

This step of the value chain brings the information out of the data, and belongs to the information part of the pyramid. For ease of graphics design, the pyramid is drawn horizontally in the value chain.

Value increases with the capability of colleting, loading, and preparing the data. There are different kinds of data types in big data, and capability of reading all types of data increases value. The data preparing capability also increases value. Typically data needs to be stored in this phase, but if real-time analysis is required, data might be stored after the actual analysis.

The preprocessing step of the value chain reflects the ETL (Extract-Transform-Load) process. This study will also consider few ETL tools as data preprocessing tool. This is why a clear understanding on the ETL process is important.

Many organizations typically use the traditional ETL tools for their structured data preprocessing purposes. The goals of ETL process are to (Simitsis, 2003):

(i) Identify the relevant information in the source side
(ii) Extract the information
(iii) Customization and integration of the information coming from multiple sources
(iv) Clean the data on the basis of requirements
(v) Propagate the data to the data ware house

The ETL process achieves these goals by three simple steps called Extract, Load, and Transform, hence the name ETL.

- **Extract:**

  Extract is the first step of ETL process which covers the data extraction from the source system and makes it accessible for further processing. The goal of this step is to retrieve required data from all the sources with little resources, and not to affect the process in terms of performance, response time negatively. Data extraction can be performed in several ways like update notification, incremental extraction and full extraction (Anon., 2013).

- **Transform:**

Transform step cleans the data, which is important to ensure the quality of the data. When the data is cleaned then transform step applies a set or rules to transform the data from source to target. This includes several tasks, such as translating coded values, encoding free-from values, sorting, joining the data from multiple sources, aggregation and splitting according to the application requirements.

- **Load:**

The load phase loads the transformed data into the end target. Depending on the requirements of the applications, this process varies widely. Typically the target of the load phase is the databases or data warehouses. During the load step it is also necessary to ensure that the load is performed correctly with the minimal resources usage.

The ETL process framework is shown in the Figure 11 below.

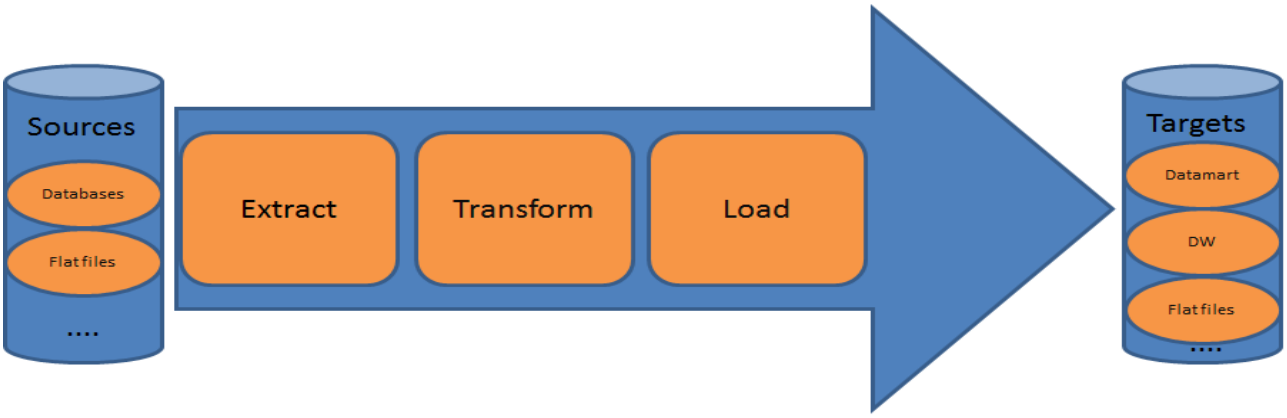

Figure 11: Typical ETL process framework

Few tools, such as Oracle Data integrator perform Extract-Load-Transform (ELT) to enhance efficiency.

2.3.3. Processing and Visualization

This step of the value chain creates the highest value. This step can also be called as 'Analytics and Visualization'. This step lies into both knowledge and wisdom parts of the pyramid. Descriptive analysis works on past and present results and answers questions, such as what happened, what happens, and how it went. On the other hand, diagnostic and predictive analysis investigate the results and answer questions, such as why it happened, who did and what is going to happen. Both the processes increase value and bring knowledge. Prescriptive analysis works on future and

includes questions, such as what is needed to be done and why, also brings wisdom. Wisdom has the highest value in the value chain (Stein, 2012).

The big data value chain shows the process of converting data into information, knowledge, and finally to wisdom. It also shows that the data preprocessing plays an important mediator role to enable information and generates insights from data.

A typical flow diagram for managing big data focusing telecom industry is presented below. From the Figure 12 it is clear that, the flow diagram follows the value chain.

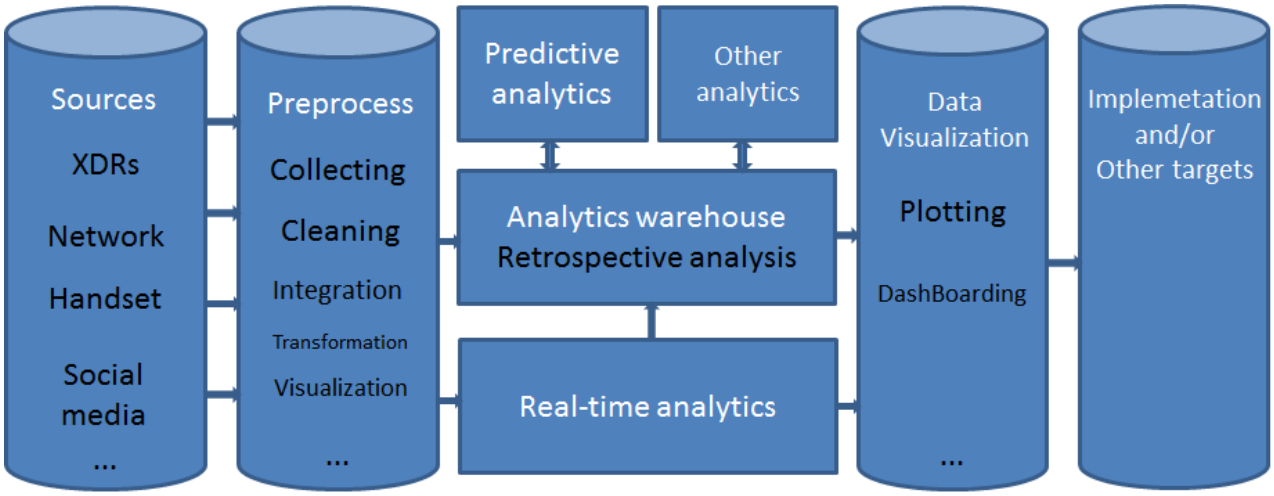

Figure 12: Typical Flow diagram for managing big data focusing telecom industry

Analytics warehouse is the process phase where typically all the analytics take places, such as retrospective analysis. Another important (specially focusing the velocity characteristics of big data) analytics type is Real-time analytics.

Real-Time Analytics:

Real-time data analytic refers to the analytics that are able to be accessed as they come into the system. Real-time analytics require the ability to process and analyze parallel streams of data as they come in from the network or from other sources, before they are ever stored in a database.

Table 3 shows how telecom operators can gain advantages by implementing real-time analytics infrastructure instead of legacy analytics infrastructure.

Criteria	Legacy Analytics Infrastructure	Real-time analytics Infrastructure
Storage cost	High	Low
Analytics	Offline	Real-time
Data loading speed	Low	High
Data loading time	Long	Average 50 percent faster
Administration time	Long	Average 60 percent faster
Complex query response time	Hours/Days	Minutes/Seconds
Data Compression technique	Not matured	Average 40 to 50 percent more
Support Cost	High	Low

Table 3: Legacy analytics infrastructure Vs. Real-time analytics infrastructure **(**Banerjee**,** 2011**)**

Figure 13 is a survey result, where 65 global operators participated, showing the potential usages of real-time analytics. In the survey 65% of the participants pointed that real-time analytics can be used for operational planning, where 62% of them pointed that the best use of real-time analytics is real-time service assurance. Real-time analytics can also be useful for other analyses, such as price and product mix optimization, conduct advanced analytics and network optimization.

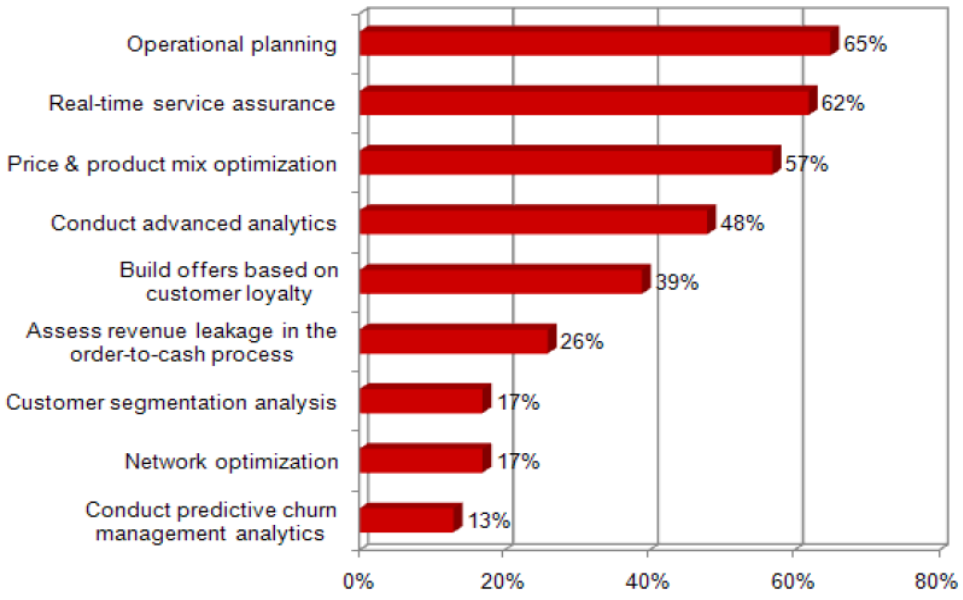

Figure 13: Use of real-time analytics to deliver on defined business objectives for operators **(**Banerjee**,** 2011**)**

Above discussion, figure, and table implies, that the real-time data analytics can play vital role to meet the business objectives for a network service provider company.

Different organizations are currently offering big data technologies and tools for different steps of the value chain. A big data landscape is presented in Figure 20, Appendix 1A.

2.4.*Big data Challenges*

Big data also has some significant challenges, some of them are mentioned below:

- **Storage**: The first and foremost challenge of big data is the storing. Traditional data ware houses are not typically made for it. Organizations that are trying to adopt big data strategy need to build a new warehouse, which is capable of storing big data for them.
- **Complexity**: The three dimensions of big data, namely volume, velocity, and variety make it more complex and challenging to analyze than the other traditional data.
- **Management**: Big data management systems available in the current market are not able to satisfy the needs of it (Ji, et al., 2012), thus a re-construction of the information framework is needed. Re-organizing the data in this re-constructed frame work is another big challenge.
- **Preprocessing**: Finding out the right data from big amount of data which also have verities in it, is typically challenging. Big data preprocessing requires collection capability, statistical analysis, and integration capability of large amount of data. Traditional extract-transform-load (ETL) tools are not able to fulfill these requirements.
- **Analytics**: Big data analytic is a highly mathematics intensive analytic modeling exercise which requires proper tools and skilled people. Big data also requires highly capable tools for data visualization, because traditional tools are typically made for small amount of data.
- **Utilization gap**: Christine Moorman stated that the biggest challenge regarding big data is the Utilization gap (Moorman, 2013). When asked to report the percentage of project in which their companies are using marketing analytics that are available, CMOs report a dismal of only 30% usage rate (Moorman, 2013). In (McGuire, 2013), the hardest challenge of big data is mentioned as, taking the insights generated from the analytics and utilizing them to change the way business operates.
- **Lack of skilled people**: As big data is a new concept and requires newer technologies; there is a lack of skilled people for it. According to Gartner, big data demand will reach around 4.4 million jobs globally by 2015, with two third of these positions remaining unfilled (Gartner, 2012).
- **Privacy**: In several research studies, privacy concern has defined as the biggest barrier for big data (Kaisler, et al., 2013; Smith, et al., 2012; Demchenko, et al., 2012). When it comes to the customer personal data and how it is used, people generally don't like surprises. The study (Smith, et al., 2012) shows, how the location data and the social media data are hampering users' privacy, and the users are not concerned about it. The social media data is

being one big topic about the users' privacy issue in recent days but the user location data being as a privacy issue has not got that much attention yet (Smith, et al., 2012).

- **Security**: Big data security management is also one challenging task. Traditional security mechanisms, which are tailored to secure the small-scale data, are inadequate for it.
- **Real-time analysis**: Big data requires real-time analysis, which is sometimes challenging. Real-time analysis requires high-velocity streaming analysis of big amount of data and typical data analysis tools are incapable of doing so.
- **Additional challenges**: There are more additional challenges regarding big data, such as transportation of data, dynamic design requirement, and scaling.

An organization, before starting big data projects needs to make new policies to mitigate these challenges, and select tools which are truly capable for it. Otherwise, the project acquires a large possibility to be failed in the middle of the process which will cause the organization financial loss.

2.5. *Role of Hadoop in big data*

Traditional databases are good at very rapid, interactive queries on moderate and small datasets, but typically run out of stem and can take long time, once the data starts to get very large. On the other hand, Hadoop with its parallel processing is poor at small queries, but is perfect for larger workloads that take large amount of data and more complex queries (Metascale, 2013).

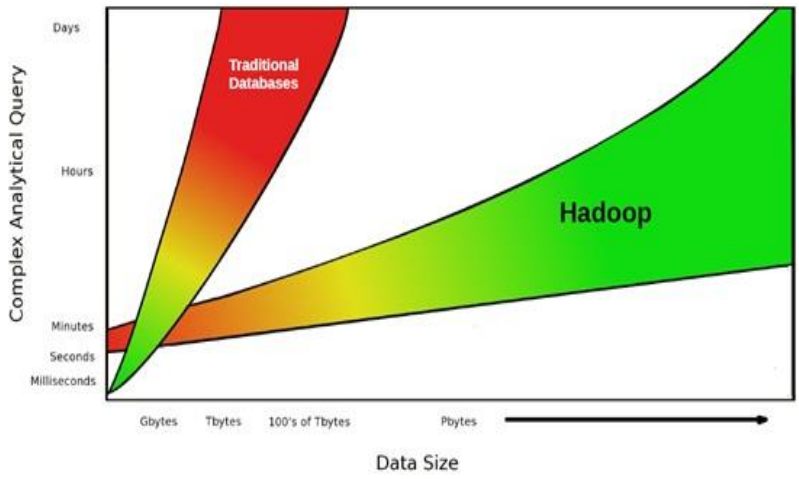

Figure 14: Capability comparison of Hadoop with traditional databases **(**Metascale**,** 2013**)**

Because of Hadoop's capability of analyzing and storing big amount of data, it is often seen as a synonym for big data. This is why this thesis includes Hadoop description in brief in this chapter.

2.5.1. Hadoop Characteristics

The followings best characterize Hadoop:

- Hadoop is open source.
- Hadoop is scalable. This is one important characteristic of Hadoop that makes it suitable for big data analytics. It allows new nodes to be added as needed, without concerning the data formats or how data is loaded.
- Hadoop is flexible, schema-less, and can absorb any types or formats of data. Where traditional databases typically fail to load unstructured data.
- Hadoop is fault tolerable. Whenever a node got disabled the systems redirects the process to another location of the data and keep the process running.
- Hadoop is fast. It is also able to run parallel processing as big and the background batch jobs in the same server at a time. This saves the user from acquiring additional hardware for a database system to process the data.

2.5.2. Hadoop Architecture

Hadoop architecture consists two main layers, Hadoop distributed file system (HDFS) and MapReduce.

Hadoop distribute file system (HDFS): HDFS is a distributed, scalable and portable file system written in Java. HDFS typically contains one name node and a cluster of data nodes. Each data node serves up blocks of data over the network using a block protocol specifically to HDFS. Each node does not require a data node to be present. HDFS replicates the data across multiple hosts and hence does not require Redundant Array of Independent Disks (RAID) to store, which makes it highly reliable. The term replication value represents the number of nodes the files are stored in. HDFS uses default replication value 3 (Wikipedia, 2013). The data nodes are capable of communicating with each other to rebalance the data, to move copies around and to keep the replication of data high.

MapReduce: MapReduce is a programming model for processing large data sets with a parallel and distributed algorithm on a cluster. In MapReduce, map performs filtering and storing, and reduce performs a summary operation. In map step master node takes the input or the problem, it then divides the problem into smaller sub-problems and distributes them to the worker node. A worker node also has the capability to divide the sub-problems into several sub-sub-problems and distribute

them to other worker nodes under it, which leads to a multi-level tree structure. The worker node processes the smaller problems and passes the answer back to its master node. If all the mapping operations are independent, all maps can be performed in parallel. The same for the reduce parts as well, allowing a distributed parallel processing which can save large amounts of time of the users.

Over a period of time to make things simpler, user friendly, and efficient, several other products have been developed around the Hadoop ecosystem.

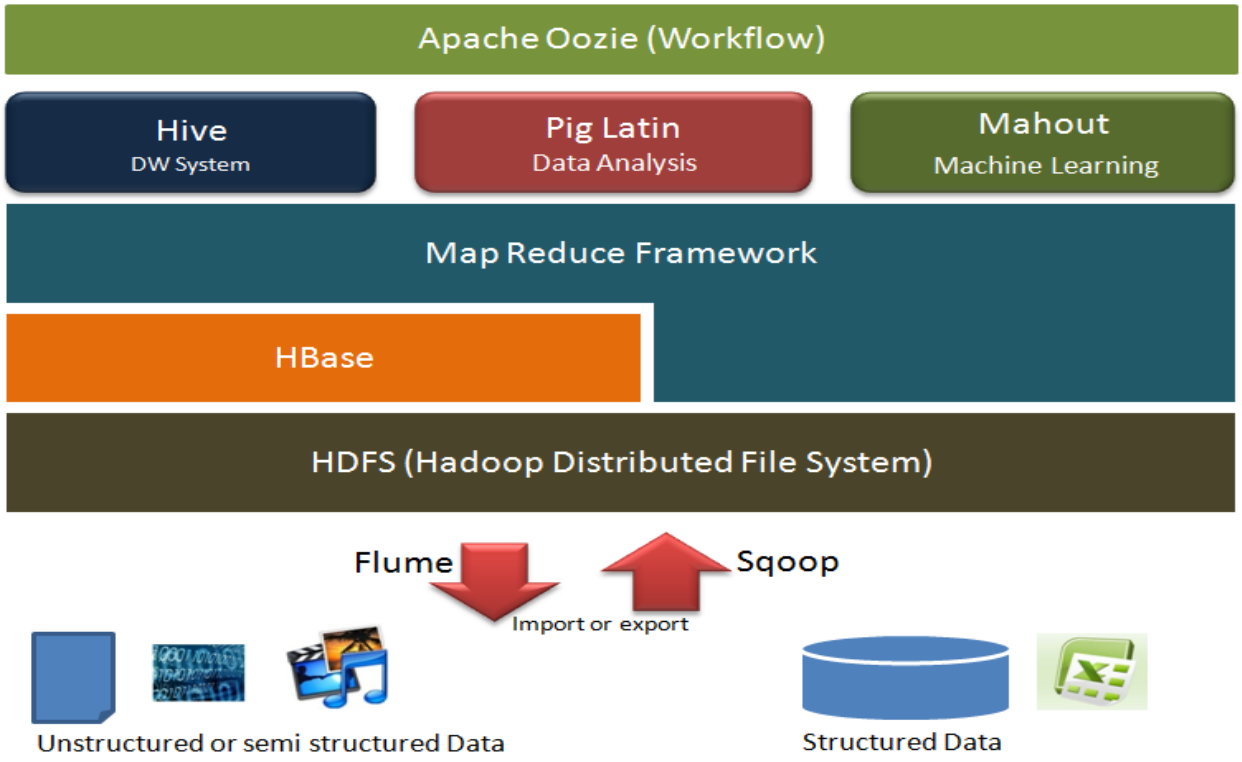

Figure 15: Hadoop Ecosystem **(**Pradhan**,** 2012**)**

Figure 15 shows the current Hadoop ecosystem. Flume is an Apache product which is a distributed, reliable and available service for efficiently collecting, aggregating, and moving large amounts of unstructured data. Sqoop designed for efficiently transferring bulk data between HDFS and relational databases. HBase is distributed, scalable and big data store which is also known as Hadoop database. Apache Pig is an analyzing platform for large data sets. Apache Pig has high level language for expressing data analysis programs. Hive is a data warehouse system for Hadoop. Apache Mahout is a machine learning library which was built to fulfill the goal of building a scalable machine learning libraries. Apache Oozie is the workflow scheduler for Hadoop, which is responsible for managing Apache Hadoop jobs (Apache, 2012).

2.5.3. Hadoop's limitations

Hadoop also has some limitations. The limitations are:

- With default replication value 3, Hadoop saves the same data set in three different places. That is why even small data sometimes become big data in Hadoop.
- Hadoop has a very limited SQL support (actian, 2012). Lots of companies are already expert on the SQL, and limited SQL support makes Hadoop a bad choice for some of them.
- Few experts find MapReduce a challenging framework and the machine learning library Mahout difficult to implement.
- Hadoop is not able to do real-time analytics. MapReduce is a batch-based architecture, that means it does not lend itself to use cases which need real-time data access. But recently, with some new tools, such as Storm, Cassandra, and Mongo, Hadoop is getting capable of doing real-time analysis.
- Hadoop is not practically a simple technology. Linux and Java skills are critical for making its environment. DBAs will need to learn new skills before they can adopt Hadoop tools. Most importantly, it is also not easy to connect it to legacy systems.

Distributed computing system of Hadoop is a great promise for handling large data, but it lacks the toolset that are familiar with on a single machine (Prekopcsak, et al., 2013). These are why Hadoop is not used in actual tests for this thesis. Many data analysis tools now have Hadoop extensions, which connects them to Hadoop data bases. In this thesis, Hadoop extension will be considered as an important feature while choosing the preprocessing tools.

As a conclusion, big data can be considered as an asset to the organizations. Utilization of proper tools can facilitate the organizations, such as telecom industry getting benefits from their big data in different ways.

3. Big data in Telecom Industry

3.1. *Telecom operators' data volume growth*

Currently there are little more than 6.2 billion mobile subscriptions worldwide, and this number is predicted to reach 9 billion by 2017 (Ericsson, 2012). With large subscriber base, telecom operators typically need to handle big amount of subscriber data. In addition, every call, internet connection and sending of SMS generates network data for operators. After introduction of smart phones, YouTube, Facebook and possibility of watching TV from the internet, data traffic in operators' network has increases heavily. According to (Cisco, 2013), monthly global mobile data traffic will surpass 10 exabytes in 2017. The increment of networked devices and applications means more data is being collected than ever before.

In 2011, global mobile data traffic was eight times greater than the total global internet traffic in 2000. More than 50 percent of Facebook users are mobile users, staggering 488 million mobile users (Aginsky, 2012). According to Gartner, 1.8 billion mobile phones were sold and among those 31 percent were smartphones in 2011 (Egham, 2013). According to the Cisco VNI Mobile Traffic Forecast, the typical smartphone generated 35 times mobile data traffic which is around 150MB per month than the typical basic feature cell phone in 2011 (Cisco, 2013).

In 2012, Global mobile data traffic grew 70 percent in 2012 reaching 885 petabytes per month at the end of 2012. Mobile video traffic has exceeded 50 percent for the first time in 2012 and the mobile network connection speeds also been doubled in 2012. Though smartphones were only 18 percent of total global handsets, still 92 percent of total global handset traffic was represented by smartphones (Cisco, 2013). Increasing number of smartphones is going to affect the global network traffic vastly.

Device type	Growth in devices, 2012-2017 CAGR (in percentage)	Growth in mobile data traffic 2012-2017 CAGR (in percentage)
Smartphone	20	81
Tablet	46	113
Laptop	11	31
M2M Module	36	89

Table 4: Percentage Comparison table of global device unit growth and global mobile data traffic growth **(**Cisco**,** 2013**)**

With this increment of smart devices usage, consumers are utilizing the network for several purposes. In Appendix 2A, Figure 21 demonstrates the expected situation in 2015 regarding certain types of traffic in the network, especially from an individual user's point of view.

In (Cisco, 2013), a forecast of mobile data traffic growth by the year 2017 has been presented and it is as big as 11.2 exabytes per month.

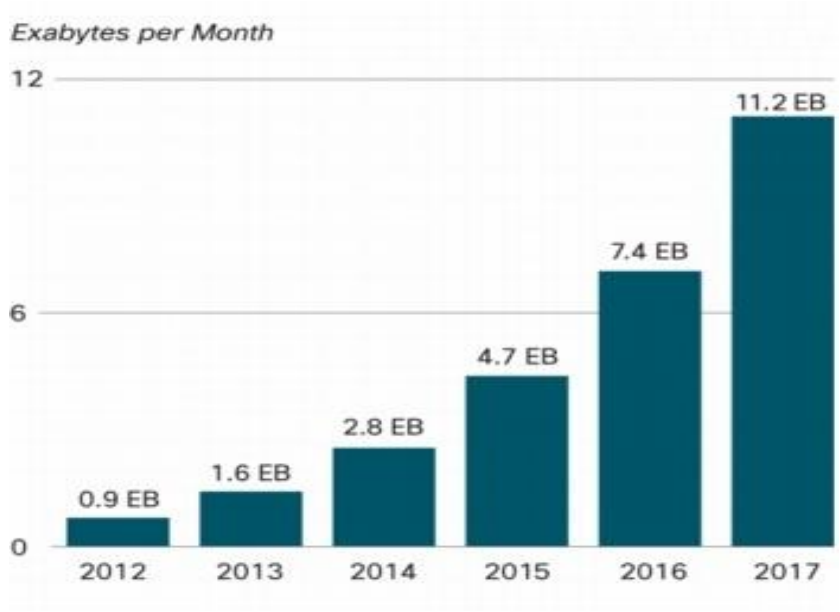

Figure 16: Cisco forecast on mobile data traffic growth by 2017 **(**Cisco**,** 2013**)**

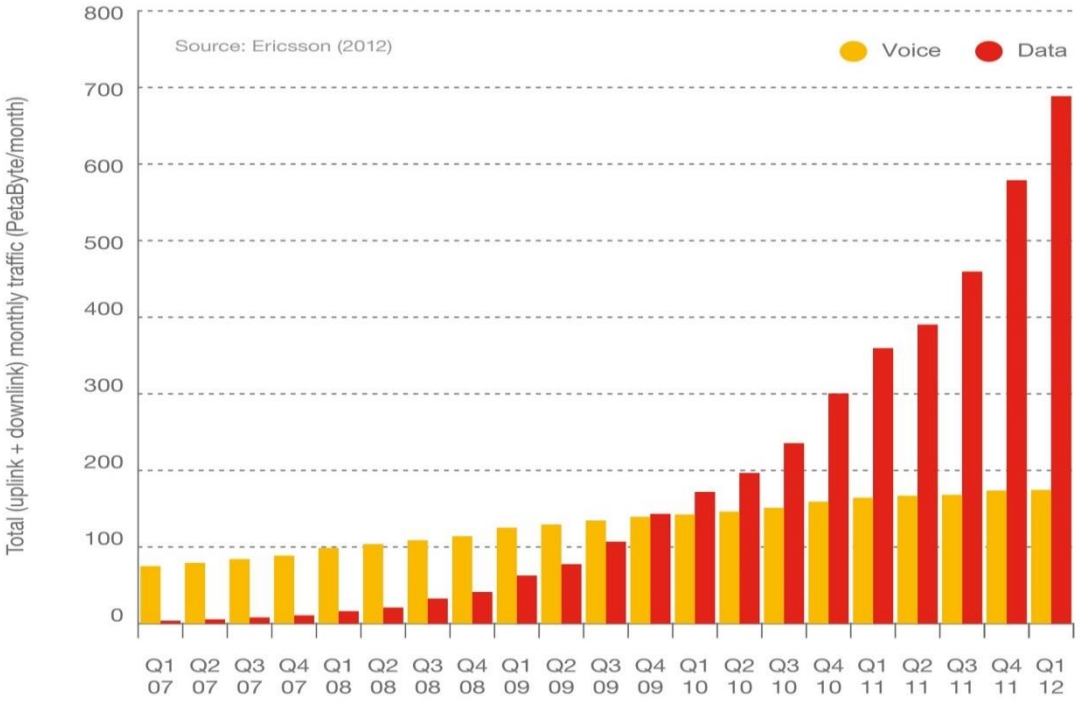

Figure 17: Global total data traffic in mobile networks, 2007-2012 **(**Ericsson**,** 2012**)**

Figure 17 shows, how the amount of data traffic is growing compared with voice traffic in the mobile networks. This also shows that the mobile operator companies are becoming data service

provider from only voice service provider. This is making the mobile operators save more and more data.

3.1.1. Data types and Data sources

In telecom industry most of the early big data efforts are targeted at analyzing the existing data such as CDRs, network data and subscriber data. More than half of the operator experts define internal data as the primary source of big data within their organizations (IBM, 2012).

There are three types of data in telecommunications industry form an operator's point of view, namely Subscriber data, Network data and Call detail record (Weiss, 2005).

Subscriber data

Telecommunication operators typically have millions of subscribers. By necessity this requires maintaining an information database of these subscribers. For example, operator saves the subscriber account information which includes e.g. name and address, billing information including payment methods and details of payment. Subscriber data also includes the information about the connection, disconnection-reconnection and itemized information of the services the subscriber uses. An operator also saves traffic data, such as information identifying the sender and recipient, routing information and online tracing of the communications of an individual subscriber. Subscriber data contains structured, semi-structured, and unstructured data. Subscriber profile data originates from the network systems such as Home Location Register (HLR) or Customer relationship management (CRM).

Network data

Network data is the data generated by the network elements. Nearly all equipment of the telecom network is capable of generating error and status messages, which lead to a large amount of network data. These data contain the timestamp, a string that uniquely identifies the hardware or software component generating the message, and code that explains why the message is being generated. Network data also contains both structured and unstructured data. The alarms or error codes can be categorized as structured data, but the error message if not predefined, is unstructured data.

C2S	S2C	Short Description	Unit	Long Description	Protocol
1		L4 Proto	1/2	1 = TCP, 2 = UDP	All
2	38	Protocol	3/4	3 = RTP, 4 = RTCP	All
3	39	IP address	-	Client/Server IP addresses	All
4	40	L4 port	-	TCP/UDP port addresses for the Client/Server	All
5	41	Internal	0/1	1 = internal ip	All
6	42	Packets	-	Number of packets Tstat has seen belonging to the flow	All
7	43	IPG	ms	Inter Packet Gap (IPG)	All
8	44	Jitter AVG	ms/ts	Jitter (average): - if RTP, computed by Tstat as in RFC3550 [ms] - if RTCP, extracted from the RTCP header [codec timestamps units]; - if TCP, computed using only data packets [ms]	All
9	45	Jitter Max	ms/ts	Jitter (max) - if RTP, computed by Tstat as in RFC3550 [ms] - if RTCP, extracted from the RTCP header [codec timestamps units] - if TCP, computed using only data packets [ms]	All
10	46	Jitter Min	ms/ts	Jitter (min) - if RTP, computed by Tstat as in RFC3550 [ms] - if RTCP, extracted from the RTCP header [codec timestamps units] - if TCP, computed using only data packets [ms]	All

Table 5: Typical network log file **(**Tstat**,** n.d.)

Call detail record

Every time a call is placed on the telecommunications network, descriptive information about the call is saved as Call Detail Record (CDR). The number of call detail records that are generated and stored in an operator's database is large. For example, AT&T (an American multinational telecommunications corporation) long distance customers alone generate over 300 million CDRs per day (Weiss, 2005).

Option	Value/Example	Description
accountcode	12345	account ID
src	12565551212	The calling party's caller ID number
dst	102	The destination extension for the call
dcontext	PublicExtensions	The destination context for the call
clid	"Big Bird" <12565551212>	The full caller ID, including the name, of the calling party.
channel	SIP/0004F2040808-a1bc23ef	The calling party's channel
dstchannel	SIP/0004F2046969-9786b0b0	The called party's channel
lastapp	Dial	The last dial plan application that was executed
lastdata	SIP/0004F2046969,30,tT	The arguments passed to the last app
start	26.10.2010 12:00	The start time of the call
answer	26.10.2010 12:00	The answered time of the call
end	26.10.2010 12:03	The end time of the call
duration	195	The number of seconds between the start and end times for the call
billsec	180	The number of seconds between the answer and end times for the call
disposition	ANSWERED	An indication of what happened to the call
amaflags	DOCUMENTATION	The Automatic Message Accounting (AMA) flag associated with this call
userfield	PerMinuteCharge:0.02	A general-purpose user field
uniqueid	1288112400.1	The unique ID for the src channel

Table 6: Typical CDR Fields (Madsen, et al., n.d.)

With the emerging smartphone usage, operators have come up with another transaction record naming Extended Data Record (XDRs). XDRs capture other transaction records, e.g. purchase history, download and upload history and recharge or payment history.

Telecom operators also store other business data, such as marketing data, billing data, product and service data.

Typical data sources

According to IBM and Said Business School of University of Oxford research, the big data sources for operators are the i) Transactions, ii) Log data, iii) Phone calls, iv) Events, v) Social media, vi) Geospatial, vii) E-mails, viii) External feeds, ix) Free-from text, and x) Sensors (IBM, 2012).

This study describes the typical data sources for operators as follow.

- **Subscribers** are one data source for telecom operators. Telecom operators typically stores petabytes of subscriber data.

- **Network usage** is another data source which generates subscribers' network usage data, such as CDRs, XDRs, Internet Protocol Detail Record (IPDR) and data from Business Support Systems (BSS) or Operational Support Systems (OSS).

- **Network Elements** are generating network data.

- Introduction of IoT and M2M is going to increase the number of **connected devices** to 24 billion by 2020 and most of them are going to use mobile internet services through mobile operators' network which means lots of data growth for the operators (Malik, 2011). For example, (Karikoski, 2012) presents example of data collected form users' mobile devices for research purpose.

Type	Details
Phone calls	Duration in/out, type (in/out/missed)
SMS, MMS	No. of messages sent/received, length
Application	Foreground usage, installation, background processes
Browsing	HTTP traffic
Bluetooth scan	Names, MAC addresses
WiFi scan	Names, MAC Addresses
Location	MCC, MNC, LAC and cell ID
Network session	Bearer, IAP, uploads, downloads
Energy	Battery and charging status

Table 7: Data types collected by MobiTrack from handset **(**Karikoski**,** 2012**)**

- **Products & services**. When operators offer some services they also create an opportunity to collect the data about the services to analyze e.g. service performance, subscribers' behavior and target pattern.

- In the era of smartphones, **social media** is creating large amounts of data for the operators. Subscribers are using the mobile network to browse their Facebook, LinkedIn, Yahoo or Google profiles form where data can be collected by the operators.

- **Other External sources,** such as transportation services and financial services.

In (Acker, et al., 2013), data sources for operators are summarized as follow:

Network	Product	Marketing and Sales	Customer Care	Billing
Network events data	Product catalog	Customer device data	Order data	CDRs
CDRs	Product life-cycle data	Sales channel data	Fault handling data -Contract data -Problem type -Resolution time and rates -Repeated faults	Traffic data
SMS and MMS data	Product and platform costs data	ARPU classification		Usage history data
Volume of data traffic	Product usage data	Marketing response data		Customer account data
Probes data	Product delivery management data	Segmentation data		
Handset data		Usage pattern	Call center logs	
Technical fault data			Termination reasons	

Table 8: Potential data sources and available data for operators (Acker, et al., 2013)

Table 8 shows the typical potential data availabilities in different departments of a telecom operator.

3.2. *Big data use case domains for operators*

A table containing potential big data use cases for telecom industry was built for this research by using conducted interviews and study. In Table 22 (Appendix 3A), the potential use cases are presented by short names and descriptions. The use cases are divided into different use case domains. The use case domains represent the business areas where the use cases are focused on. Possible required data along with data sources for each use cases are also mentioned. The requirements to meet these use cases are also presented in a different column.

This thesis covers two types of use cases for operators in different domains, namely external use cases and internal use cases. Internal use cases e.g. increase ARPU, reduce churn and network improvement represent the use cases which typically require internal data. On the other hand, external use cases represent the use case which typically requires data from external sources along with internal data. External use will be to generate revenue form big data, such as by selling insights to various verticals, e.g. Verizon sell insights Verizon (Precision market insights) and Telefonica (Smart Steps) (ITU, 2013).

The key use case domains are [from Appendix 3A]:

- **Improvement of different characteristics of the Network:** Operators have always been concerned about network performance improvement. Now with big data analytics operators

can identify problems, perform real-time troubleshooting and fix network performance issues. This might offer improved network quality, lower operating cost and higher customer satisfaction.

For example, Turkcell (leading mobile phone operators of Turkey) has rolled out a big data application, to help it recover from network failures by correlating data from different sources in real time to identify the root cause of failure (Middleton, 2013).

All SON (Self-organizing network) automation such as provisioning, configuring and commissioning can be adapted to changes in the environment and traffic demand based on the insights gained from big data analytics (Ericsson, 2013). It can also let the operators to prioritize the alarms, which would be very useful to save time and service failure (Wallin & Landen, 2008).

The typical data types required are network element data, CDRs, location data, XDRs, traffic data and network events data.

These use cases typically require capability of collecting data from the diverse sources and aggregation, capability of real-time data analysis and predictive analysis.

Typical challenge to meet these use cases is real-time analytics, needed for e.g. real-time network optimization and real-time network monitoring. This includes high-velocity streams of CDRs to detect critical network events as they happen.

- **Marketing and Sales:** Marketing and sales can be considered as the largest domain of big data usage in telecom industry. Proper big data analytics allow the operators to create more intelligent marketing campaigns, and to do sales analytics to increase the sales. It can also be utilized to improve the results of marketing promotions, increase revenue and help to prevent the customer churn.

  For example, Globe Telecom (Telecommunications Company in the Philippines) uses big data analytics to improve effectiveness of promotions by 600% (IBM, 2010).

  The typical required data for these use cases are subscriber location data, subscriber data, social media data and previous campaigns data.

  These use cases will typically require capability of collecting data from diverse sources, big data storage and data correlating capability. Other typical requirements are capability of unstructured data analysis and capability of sentiment analysis. Delivering these types of intelligent marketing campaigns requires rapidly processing of high volumes of location data, automatically combining it with other data, and making it deliverable in real time.

  One big challenge to meet these use cases is unstructured data (e.g. text, images, and videos) analysis, which requires advanced tools and skilled personnel.

- **Security:** Big data analytic allows telecommunication companies to detect and analyze high-velocity fraud activities in real time and take actions immediately. It can also be utilized for e.g. real time cyber security monitoring, information security management and preventing unauthorized physical or virtual access.

  For example, Elisa Oyj (Finnish Telecommunications Company) offering service called Elisa Vahti, which provides real-time security monitoring of subscribers' home/cottage/office (Elisa, 2013).

  By comparing subscribers' current calling behavior with a profile of his past usage, and using deviation detection and anomaly detection techniques, operators can identify super imposition frauds (Weiss, 2005). Operators can also provide security services to other organizations utilizing their big data analytics capabilities.

  These types of use cases typically require e.g. location data, XDRs and subscriber data.

  Real time data analysis capabilities, data aggregation capability and capability of correlation analysis are typical functional requirements to meet these use cases.

- **Improving customer care services:** Operators can use big data analytics to enhance the customer care services. Operators can get a good understanding on why the subscribers call, and impose an automated procedure to resolve their calls faster (Lande, 2013). Delivering a higher level of customer care can be a key strategy in differentiating an operator's brand form its competitors. Big data analytics can allow the operators to reduce the numbers of customer care calls and earn customer satisfaction by solving the subscriber problem in real-time.

  For example, one Tier 1 mobile service provider in United States transforms call centers with real-time access to customer and product data (IBM, 2012).

  Improving the customer care service typically requires subscriber data, network performance data, network events data, customer care agents' data and historical data.

  Capability of data aggregation, real-time dash boarding, and access to diverse data sources are the typical functional requirements for use cases in this domain.

  These use cases require large volumes of data aggregation in real-time and the velocity of the data is also high, which make it challenging.

- **Business development and innovating new business models:** Big data analytics can improve operators' current business by suggesting optimization in the business strategy, such as setting new business targets, and new business models. This domain includes both external internal use cases. External use cases involve use cases, such as deliver remote cloud services to other companies and creation of profitable new business partnerships with

other companies. There are many more big data use cases in the business domain, which can improve operators' business e.g. generate own application services.

Use cases in this domain require both internal and external data. Location data, subscriber sentiment data from social media, and external data from the other partner companies such as content providers are the typical required data types.

Capability of collecting data from different sources, capability predictive analysis, capability of real-time analysis and capability of data correlating and aggregation are the typical functional requirements to meet the use cases in this domain.

Typical challenges to meet the use cases of this domain are that, they involve accessing diverse data sources and correlating them in real-time.

- **Products and Services development:** Operators can utilize big data analytics for their products and service development. They can monitor the performance analysis, margin analysis, pricing impact and stimulation, and impact of supply chain utilizing it. Historical sales data analysis of previous products and services can allow the operators to predict the possible outcome or revenue possibility of the new product or service. It also allows the operators to find out the next best product or services i.e. personalized services, according to the subscribers' usage behavior, interest and sentiment. Operators can also predict and analyze the pricing impact of any specific service or product and stimulate it if required. It can also help the operators to improve supply chain management by e.g. monitoring the performance of each partner in the supply chain, and do what-if analysis for a new product launch.

  These types of use cases typically require data from the product and service data bases, billing data and pricing data etc.

  Capability to do predictive analysis, capability to aggregate data form different sources, capability to analysis unstructured data are typical functional requirements for these use cases.

- **Billing:** With proper big data analysis, operators can ensure accurate billing for the subscribers' usage. Accurate billing assurance, bill shock prevention, billing information provisioning in real-time, are the typical use cases in this domain.

  These use cases typically require billing data, subscriber profile data, XDRs, and CDRs.

  Capability of real-time data analysis, correlation analysis, big data storage, and retrospective analysis are the typical functional requirements to meet these types of use cases.

- **Improving Transportation Service:** Operators with their big data analytics can help the government or individual subscriber to improve their transportation service. These use cases

can also be considered as external use cases. Operators can visualize the traffic situation in a specific area and let the subscribers know about it, when they require. They can also provide some pop-up services which will send the subscribers traffic information e.g. if there is road blocking on their way. Operators can also offer services such as vehicle tracking with GPS services for security purposes, route mapping, suggesting the nearest gas station when gas is running low and tracking of driven kilometers for taxation purposes.

These types of use cases will typically require data types, such as location data from the handset, and mapping data from the data bases.

Capability of real-time data collection from diverse sources and analysis, capability of data manipulation and data aggregation are the typical functional requirements for these use cases.

- **Public sector:** Operators can utilize their big data analytics in different public sector use cases. These can also be categorized as external use cases. Operators can offer services, such as locating tracking for elderly people or a device, and cars. They can also offer services like power grid information, such as information about load shedding, or information about some individual's connection bills. One important use cases would be the response to the calamities.

  Typically they require location data, weather data, and required information from other sources, such as power grid.

  The system typically needs to be capable of collecting data from diverse sources, predictive analysis and real-time analysis to meet these use cases.

- **Governmental:** Operators can help the government with their big data analytics capability in many ways, e.g. improving services, such as transportation services, healthcare services, emergency services, and educational services. Operators can also help government to offer services such as online taxation, ticket payments, transportation information provisioning.

  These use cases requires both external and internal data sources.

  Big data storing capability, diverse data types and sources access capability, real-time and predictive analysis capability are the typical functional requirements for these use cases.

- **Healthcare:** Operators can utilize their big data analytics capability to improve the healthcare services. Operators with collaboration with healthcare centers can offer services such as remote healthcare monitoring, connected hospitals, case conferencing, chronic disease management, online medical library to the subscribers. Operators can also offer services like drug authentication to the subscribers and emergency alerting and monitor to the healthcare providers. One telecom operator from France, Orange S.A. along with some

health care centers are already offering healthcare services such as connected hospitals, case conferencing, shared medical imaging, drug authentication (Orange (telecommunications), 2013).

These types of services can also gain customer satisfaction and grow number of customers, which can grow operators' businesses.

Capability of data collection from diverse sources, data correlation analysis and capability of real-time data analysis are typical functional requirements to implement the use cases under this domain.

- **Media and Entertainment:** Telecom operators with their big data analytics capability can learn subscribers' interest and sentiment about media and entertainment. Operators can utilize those learning to suggest programs and news to the subscribers accordingly. This might help the operator earning customer satisfaction and also invent some new business models, such as collaboration with media companies.

  For example, TeliaSonera (mobile network operator) offers its phone and internet subscribers subsidized access to the popular music service Spotify (TeliaSonera, 2013).

  Sentiment data from social media, usage data from devices, location data and external data from the media and entertainment organizations are the typical data types required for these use cases.

  Data collection capability, big data aggregation capability, data correlation analysis capability and capability of predictive analysis are the typical functional requirements to implement this domain's use cases.

- **Others:** There are some other big data use cases for telecom operators, such as quality control, partner analysis, and cost and contribution analysis. Others use cases might also include e.g. improving banking and insurance sectors, or providing the subscribers with services like mobile banking, mobile insurance, mobile retail shopping and mobile pricing analysis of products.

  For example, Elisa Oyj (Finnish Telecommunications Company) launched an E-Wallet that supports virtual credit cards and master card pay passes (Elisa, 2012).

Telecom operators can unearth several potential use cases by effectively monetizing the increased volume, variety, and velocity of network, subscriber, and other business data. Proper data preprocessing and analysis capabilities are required for it.

4. Data Preprocessing

According to (Acker, et al., 2013), data preprocessing can be up to 80% of the total analysis work and analyzing the data, once joined, cleaned and transformed consumes just about 20%. Willian H. Inmon (known as father of data warehousing) has stated that the data extraction, cleaning and transformation comprise the majority of the work of building a data warehouse (Bellaachia, n.d.).

Data preprocessing has important impact on the data value chain and typically includes several major tasks and techniques.

4.1. *Reasons for data preprocessing*

The raw world data is usually noisy, inconsistent and incomplete due to their typically large size and their likely origin from multiple and heterogeneous sources. Low-quality data will lead to low-quality mining results. The target of the preprocessing phase is to increase the quality of data i.e. to transform the data to information and prepare the data for actual analysis.

The quality dimensions of data are presented in Table 9.

Accuracy		Completeness
Consistency		Timeliness
Believability		Value added
Interpretability		Accessibility

Table 9: Dimensions of data quality

The real world data is noisy, where the term noisy refers to having incorrect or corrupted values. Raw data is noisy typically because:

- The data collection software may have some trouble or technical problem of providing correct data.
- Human or computer error in data entry.
- Error in data transmission, technological limitation, inconsistencies in naming conventions or data codes used.
- Containing errors, or outlier values that are deviated from the expected.

Inconsistency is another common characteristic of raw data. Data inconsistency means various copies of the data no longer agree with each other. Raw data is inconsistent typically due to:

- When different and conflicting versions of the same data appear in different places.

- Data redundancy. Data redundancy occurs when the data file/database file contains redundant and unnecessary duplicate data.
- Containing discrepancies.

Incomplete data means lacking attribute values, or certain attributes of interest, and/or missing values in the dataset.

Incomplete data occurs typically because e.g.:

- Attribute of interest may not be available.
- Other data may not be included simply because it was not considered important at the time of entry.
- Relevant data may not be recorded due to a misunderstanding or equipment malfunction.

In (Famili, et al., 1997), real world data problems are divided in to three categories, namely too much data, too little data, and fractured data. Table 10 summarizes the typical real world data problems with typical reasoning.

Too much data	Too little data	Fractured data
Corrupt and noisy data	Missing attributes	Incomplete data
Feature extraction	Missing attribute values	Multiple sources of data
Irrelevant data	Small amount of data	Data from multiple levels of granularity
Numeric/symbolic data		

Table 10: Real world data problems

Data preprocessing tries to solve the problems presented in Table 10, and prepare the raw data for the actual analysis. Other than just solving these problems data preprocessing also understands the nature of the data, and performs in-depth analysis on it through required tasks and techniques.

4.2.*Major Data preprocessing tasks and techniques*

Data preprocessing needs to perform several tasks or activities in order to ensure the data quality before the analysis phase. It is not necessary that every task has to be performed for every data set.

In (Tanasa & Trousse, 2004), data preprocessing is divided into two parts, naming Classical Preprocessing and Advanced Preprocessing. The classical preprocessing can be further divided into data fusion, data cleaning and data structuration phases. The advanced data preprocessing part includes only the data summarization phase.

This study classifies the preprocessing phase on to the followings:

Data import:

In import phase data is collected, loaded, and checked. Accessing to diverse data sources, reading and loading of different types of data are important capabilities needed for big data preprocessing.

Data summarization:

The target of data summarization is to have an overall idea about the data. It identifies the typical properties or characteristics of the data, which helps the analysts to get an overview about the data. The data summarization includes two major sub-tasks, namely measuring the central tendency and dispersion of the data. The central tendency measures the mean, median, mode and midrange of the data. Data dispersion measures the quartiles, interquartile range, variance and standard deviation. There are specified statistical algorithms and analysis processes for performing these tasks. Data visualization and identifying data properties are also two important sub-tasks performed in this phase.

Data Cleaning:

Data cleaning cleans the data by filling missing values, smoothing noisy data, removing outliers and resolving inconsistencies. A proper cleaning of the data ensures the quality of the data and makes the data trustworthy.

There are numbers of algorithms for data cleaning categorized under missing value analysis, binning, regression and clustering methods.

Data Integration:

Data integration is the process of collecting and combining data from different sources into a coherent data store. Careful data integration helps reducing redundancies, inconsistencies and improving mining speed and quality.

Schema integration is one major sub-task of data integration, which integrates Meta data from different sources. Data integration also involves entity identification, which performs identifying real world entities form multiple data sources. In addition, data integration can handle redundancies in the data by using correlation analysis for instance.

Data Transformation:

Sometimes data may need to be transformed into forms appropriate for mining. Data transformation is the process to transforming the raw data to another appropriate form. Data transformation typically involves smoothing the data, data aggregation, data generalization, data normalization and attribute construction (Han, et al., 2011).

Data Reduction:

Data reduction obtains a reduced representation of the data set, which is much smaller in volume but yet can produce the same or almost same analytical results and saves time and effort. Discretization, numerosity reduction, dimensionality reduction, data cube aggregation and attribute subset selection are the typical sub-tasks performed in data reduction phase.

Data feed to analysis process:

When the data is prepared for the analysis, it needs to be converted to the format which is required, and then fed to the analysis process.

Data import	Data summarization	Data Cleaning	Data Integration	Data Transformation	Data Reduction
Collect the data	Identify data properties	Missing value analysis	Schema integration	Smoothing	Data cube aggregation
Load the data	Outlier detection	Binning	Object matching	Data aggregation	Attribute subset selection
Define data types	Measure of central tendency	Regression	Entity identification	Data generalization	Dimensionality reduction (PCA)
	Measure of data dispersion	Clustering	Correlation analysis	Data normalization	Numerosity reduction
	Data visualization			Attribute construction	Discretization

Table 11: Major data preprocessing tasks and corresponding sub-tasks

Data preprocessing also involves data visualization as a major task where the prepared data is presented or visualized to make sure the data is prepared for the actual analysis. Aside from bar charts, pie charts and line graphs which are used in most statistical data presentation, data visualization also includes multi-dimensional plotting, e.g. histograms, quartile plots, q-q plots, scatter plots and loess curves.

In (Famili, et al., 1997), data preprocessing techniques have been divided into three parts, naming Data Transformation, Information Gathering, and New Information Generation. These three techniques also classify the six major preprocessing tasks described above.

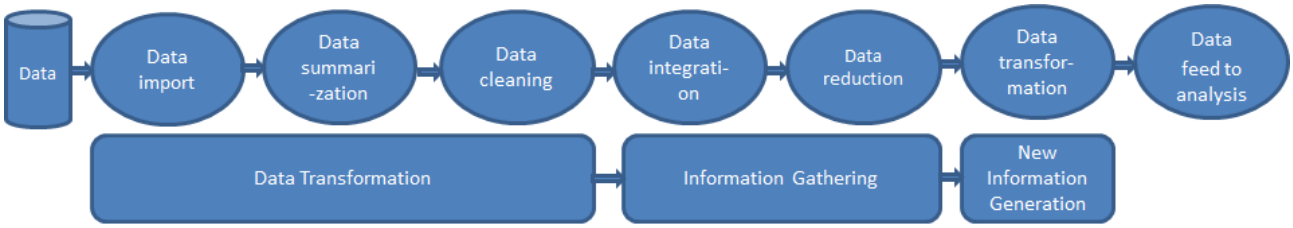

Figure 18: Major data preprocessing tasks and techniques (typical process flow)

Figure 18 represents typical data preprocessing process including major tasks and techniques.

Data Transformation:

In (Famili, et al., 1997), data transformation is mentioned as one major data preprocessing technique which covers three major preprocessing tasks, such as data import, data summarization, and data cleaning at some extent. Data transformation as a preprocessing technique typically involves data filtering, data ordering, data editing and noise modeling.

Information Gathering:

Information gathering as a technique ensures better understanding of data, by gathering and presenting the information that the data provide. If all data characteristics are not known, the analysis might not get properly guided which gives limited or incomplete result, this is why information gathering is important. Information gathering is part of data preprocessing techniques, which involves several sub-tasks, such as data visualization, data elimination, data selection, principal component analysis and data sampling. This step covers the data integration and data reduction from major preprocessing tasks.

New Information Generation:

Generation of New Information is considered as the most important part of the data preprocessing as the new information generated in this step is typically used for the analysis. It covers data transformation part from the major tasks of data preprocessing.

Data transformation	Information gathering	Generation of new information
Data import	Data visualization	Data engineering
Data filtering	Data elimination	Time series analysis
Data ordering	Data selection	Data fusion
Data editing	Principal component analysis	Simulation
Noise modeling	Data sampling	Dimensional analysis
Data summarization	Data integration	Constructive induction

Table 12: Data preprocessing techniques and corresponding sub-techniques

The data preprocessing sub-tasks and sub-techniques are described in Appendix 4A.

4.3.*Big data preprocessing challenges*

The typical big data preprocessing challenges are summarized as:

- Data problems identification: Data preprocessing is responsible to solve the real world data problems (e.g. mentioned in Table 10). In big data, identifying the data problems is challenging because of its large volume, variety and velocity.
- Proper technique selection: In big data it is challenging to select suitable preprocessing technique as the data might have several hidden problems, such as fractured data.
- Domain expert presence required: To perform meaningful data preprocessing, the domain expert needs to be a part of the preprocessing task, otherwise the domain needs to be studied extensively before starting preprocessing the data of it (Famili, et al., 1997). It is challenging to find an expert having both skills.
- Data preprocessing might be iterative: Data preprocessing might be iterative in many cases. For big data, the required numbers of iterations get larger, which is challenging and time consuming.
- Lack of tools: One big challenge in the area of research on big data preprocessing is to develop a tool box or an expert system that can provide advice for selection and use of the best technique. There are not many tools available which can perform all the preprocessing tasks and techniques on big data.
- Lack of experts: The analyst needs to properly understand the data first to specify, how to join the data, what to do with the missing values, and what aggregation technique over the data needs to be performed. It is not just a problem of hardware; it is a problem of talent (Famili, et al., 1997).

Data preprocessing involves several tasks and techniques. To perform the major preprocessing tasks and techniques, proper preprocessing tools (which are capable of performing the tasks efficiently) are needed to be selected beforehand.

5. Preprocessing Solutions

Data preprocessing requires a number of preprocessing, analytics and performance features. The tools which are capable of fulfilling the feature requirements are best suited for preprocessing solutions.

5.1. *Feature requirements*

To sort out the important data preprocessing features, questions such as ‘what are the feature requirements for a suitable preprocessing tool?’ was included in the interviews (Jony, 2013). By combining experts’ answers, the requirements in use case table (Table 22) and understanding on preprocessing, required preprocessing features were sorted out in this research.

The required features for data preprocessing are divided into three parts, namely:

1. Data preprocessing features

2. Analytics features

3. Performance and usability features

The analytics features and performance and usability features are additionally divided into four sub divisions.

5.1.1. Data preprocessing features

The typical functional requirements for data preprocessing are listed in Table 13. The data preprocessing features also reflects the major data preprocessing tasks and techniques described in Chapter 4.2. The selected tools will be evaluated according to their capability to meet these functional requirements and will be scored accordingly.

Preprocessing Features
Handle missing values
Filtering
Aggregation
Validation
Correlating
Enrichment
Data synchonization
Profiling
Outlier detection and analysis
Meta data transformation
Sampling
Discretization
Clustering
Transformation
Reduction
Name, role and value modification
Type conversion
Optimization
Attribure generation
Sorting
Rotation
Set operations
Classification
Regression
Segmentation
Manipulation
PCA

Table 13: Data preprocessing features requirements

This study has also documented the scoring criteria in Appendix 5A of the thesis. The tools are scored from 0 to 3 according to the capability of full filling the each feature requirements. For example,

Function	Scoring (0-3)	Scoring description
Missing value analysis	3	Capability to Auto detect, Impute, calculate or predict missing value
	2	Capability to fill up missing values by e.g. attribute mean, median
	1	Capability to fill up missing values by some constants like zero or infinity
	0	No capability to handle missing value

Table 14: Example: scoring levels for Missing value analysis functionality

Table 14 gives an example of scoring levels, collected from the Appendix 5A. It demonstrates, that if a tool is capable to auto detect, impute, calculate, and predict the missing values and replace them accordingly, it gets score 3 for missing value analysis functionality. If the tool is capable of filling up the missing value only with the statistical functions, it gets 2 and if it is only capable of filling up the missing value by some constants, it gets 1. The tool gets score zero, if it has no capability of missing value analysis.

5.1.2. Performance and Usability features

The performance and usability features are sub-divided into four parts namely 1) Ease of use, 2) Performance, 3) Error management, and 4) Price. The list of the features under this category is presented in the Table 15 below.

Ease of use	Performance	Error Management	Price
Self-Organizing	Low Latency	Error detection	Free
Flexibility	Availability	Auto fixing	Low price
numbers of options/features	In memory preprocess		
User friendly GUI	Smart Export and Import of data		
Discovery	Real time Preprocess		
Easy to configure	Reliable		
Easy upgrate	Accurate		
Learning	ELT support		
Efficiency	Less user intervension requirement		
Groovy Editor	Less Memory consumption		
Query language	Protect data loss		
Scripting language			
Process storage			
Process import/export			
Minimun Coding requirement			

Table 15: Performance and usability features list

The selected tools are also analyzed on the basis of these features and presented in Appendix 5C.

5.1.3. Analytics features

Analytics features in this study represent the simple analytics capabilities, which are important in data preprocessing. The features are presented in the Table 16 below. Selected tools are marked according to analytics features capabilities and presented in a table in Appendix 5D.

The analytics features are also sub-divided into four parts, namely database connectivity, advanced, text analysis, and data visualization.

Database connectivity	Advanced	Text analysis	Data Visualization
Access to all data sources	Parallel preprocessing	Signs analysis	Multi-dimensional plotting
Access to all data types	Modelling	Exploration	Dashboarding
Data extraction	Series preprocessing	Migration	Documentation
Multi-dimensional format data delivery	Real-time preprocessing	Tokenize	
NoSQL support		Rromve stop words	
Hadoop extension		Box of word creation	
High availability of Data		Steming	

Table 16: Analytics features list

5.2.*Preprocessing tools*

At first this study lists the available analytics/data mining-, ETL-, big data-specific-, reporting-, and predictive analysis-tools which are capable of fulfilling the preprocessing requirements. The tools-list is presented in Table 17, where the open source or free tools are marked as green and the commercial tools are marked as black. This list contains around fifty (50) tools, which are collected based on Internet study and the Interviews (Jony, 2013).

The tools are categorized according to their main focus and marked as 'x' mark. The ETL-tools are aimed only to do the ETL processes. The analytics/data mining-tools are mainly targeted on doing different kinds of data analysis processes such as cross validation analysis, clustering and naive bayes analysis. Many data mining and ETL tools are now capable of handling big data by implemented Hadoop extensions. The tools which are mainly focused on big data and are based on Hadoop platform are only marked as big data-tools in the list. The tools mainly suitable for the reporting and also have some data preprocessing capabilities are marked as reporting-tools. Some available tools are efficient performing predictive analysis those are marked as predictive analysis-tools in the list.

Each tool listed in the Table 17 is able to perform certain level of preprocessing tasks. The Table also presents the tools' popularity ranking (KDnuggets, 2013) and personal preprocessing qualitative ranking. The tools, which were not mentioned in (KDnuggets, 2013) are marked as NA (Not Available) in the popularity ranking. Tools, which were not ranked for personal qualitative rankings are marked as NR (Not Ranked) in Table 17.

Tools	ETL/ELT	Big data	Analytics/Data mining	Reporting	Predictive modeling	Popularity ranking	Personal qualitative ranking
KNIME			x			14	1
RapidMiner			x			1	2
Orange			x			18	3
IBM SPSS Statistics			x			10	4
R			x			2	5
Weka			x			4	6
MATLAB			x			8	7
Oracle Data Integrator	x					37	8
Talend Opend Studio			x			NA	9
Pentaho Data Integration	x					NA	10
Colver ETL	x					NA	11
TANAGRA			x			NA	12
STATISTICA ETL			x			9	13
Mathematica			x			29	14
Cloudera Standard		x				NA	15
TIBCO Spotfire				x		34	16
Microsoft Excel			x			3	17
Data preparator	x					NA	18
SAS ETL	x					15	19
Informatica PowerCenter	x					NA	20
IBM Infosphere		x				NA	NR
SAP BusinessObjects Data Integrator	x					33	NR
iWay Data Migration	x					NA	NR
Sybase ETL	x					NA	NR
Pervasive Data Integration Platform	x					NA	NR
Microsoft SQL Server			x			11	NR
Tableau				x		12	NR
Rattle			x			16	NR
JMP				x		17	NR
Gnu Octave			x			NA	NR
QlikView				x		27	NR
Salford SPM/CART/MARS/TreeNet/RF			x			28	NR
Stata			x			30	NR
KXEN					x	31	NR
Predixion software					x	22	NR
Miner 3D				x		32	NR
Bayesia				x		36	NR
Zementis		x				38	NR
XLSTAT			x			39	NR
Teradata Miner			x			40	NR
Lavastorm Analytics Engine-Public Edition			x			41	NR
WordStat			x			41	NR
Angoss			x			42	NR
Splunk Enterprise		x				NA	NR
Hortonworks Data Platform		x				NA	NR
Skytree Server		x				NA	NR
iReport				x		NA	NR
Intelsoft				x		NA	NR
BIRT project				x		NA	NR
Jaspersoft ETL	x					NA	NR
Adeptia ETL Suit	x					NA	NR
IBM SPSS Modeler			x			13	NR
BusinessObjects Data Integrator (BODI)	x					NA	NR

Table 17: Available tools having certain preprocessing capabilities with mark (x) representing their main focus, texts green represents commercial and text black represents open source tools

The Personal qualitative rankings have been achieved by:

- Online reviews
- Popularity (how many analysts are using the tool) (KDnuggets, 2013)
- Preprocessing capabilities
- Analytics, and performance and usability features
- Features descriptions from the tools' manufacturers' websites
- Preliminary testing of available tools. Preliminary testing was simple preprocessing capability testing of available tools and was not a part of final hand-on testing.

It was challenging to test the capability of each and every tool from the list and choose few of them for hands-on testing. This study minimizes the tools list to 20, based on Personal preprocessing qualitative ranking.

The tools selected for further scoring are presented in Table 18 below:

Tools	Category	Personal qualitative ranking
KNIME	Analytics/Data mining	1
RapidMiner	Analytics/Data mining	2
Orange	Analytics/Data mining	3
IBM SPSS Statistics	Analytics/Data mining	4
R	Analytics/Data mining	5
Weka	Analytics/Data mining	6
MATLAB	Analytics/Data mining	7
Oracle Data Integrator	ELT	8
Talend Open Studio	Analytics/Data mining	9
Pentaho Data Integrator	ETL	10
Clover ETL	ETL	11
TANAGRA	Analytics/Data mining	12
STATISTICA ETL	ETL	13
Mathematica	Analytics/Data mining	14
Cloudera Standard	Big data	15
TIBCO Spotfire	Reporting	16
Microsoft Excel	Analytics/Data mining	17
Data preparator	ETL	18
SAS Data Integration Studio	ETL	19
Informatica PowerCenter	ETL	20

Table 18: Selected 20 tools based on personal qualitative ranking

Among these twenty tools, all open source tools and three available commercial tools (Microsoft Excel, MATLAB and IBM SPSS Statistics) were also analyzed and scored according to their capability of data preprocessing (as described in Chapter 5.1.1) by the preliminary testing. The result is presented in the Appendix 5B. These tools were also analyzed according to their capability of fulfilling performance and usability features and analytics features; the results are presented in Appendix 5C and Appendix 5D respectively. Comparison of unavailable commercial tools is collected based on internet reviews and presented in Appendix 5E, 5F.

A tool needs to contain a number of features to be a suitable data preprocessing tool. Many available tools are capable of fulfilling the data preprocessing feature requirements. Four tools were finally selected for hands-on testing based on certain criteria. The selected tools and their performance in hands-on testing are described in the next Chapter.

6. Hands-on testing of the tools

Based on personal qualitative rankings, three open source tools and one commercial tool were chosen for hands-on testing. For the hands-on testing this study has chosen best preprocessing capable, easy to learn and less coding requiring tools. MATLAB, R and Weka are three very popular data analytics tools and also got very good scores in the preprocessing capabilities (Appendix 5B) but, these tools were not finally selected for hands-on testing. This is because, one important feature requirement for preprocessing tools for this study was 'less coding required', the target was to find out tools which can be used by "John Doe", i.e. in all areas without need to know deep level theory behind. MATLAB and R require intensive coding for data preprocessing. On the other hand, Weka performs best only on ARFF files, where it has some limitations for flat files such as CSV files. For example, CSV files cannot be read incrementally, train and test set may not be compatible if CSV files are used in Weka (Wikispaces, 2013). Therefore, Weka was also not selected for the hands-on testing.

The tools finally chosen for hands-on testing are: i. KNIME, ii. RapidMiner, iii. Orange, iv. IBM SPSS Statistics. The tools were tested through six pre-specified datasets, and four pre-specified preprocessing tasks. Tools performances were measured according to certain criteria and the results will be presented in the later section of this Chapter.

The specifications of the system used for the hands-on testing in this thesis are: 1) Processor: Intel(R) Core(TM) 2 Duo CPU E8400 @ 3.00GHz, 2) Installed memory (RAM): 4.00 GB (3.90 GB usable), 3) System type: Windows 7, 64-bit Operating System.

6.1. *Datasets and preprocessing tasks*

The pre-specified datasets are summarized as follow:

- Dataset 1 (Survey example): Data set 1 contains example survey results for e.g. 11 products/services. The columns represent the attributes, in this case which are product numbers, and the rows represent the participant IDs. All the values are numeric and represent the responses (1 to 39) of a particular participant for corresponding products.
- Dataset 2 (Mobile usage behavior): This data set is an edited version of the mobile usage behavior data set collected for research purpose by the SizzleLab. The original data set is called 'OtaSizzle Handset-based Data' collected for the project named OtaSizzle (Karikoski, 2012). OtaSizzle project used MobiTrack software for collecting data from the users' handsets and the example types of the data is presented in the Table 7 of this thesis.

The original dataset was edited by introducing lots of missing values, and few unusual values (e.g. n.a., not available) to make the required preprocessing task more advanced.

- Dataset 3 (Time series): The time series data set includes numbers of attributes as date and time, and service IDs in the rows. The dataset is an example data set of one subscriber's usage duration of a service, e.g. internet at a particular time.

- Dataset 4 (Text data): Text data represents unstructured data. This study includes one text data set for the hands-on testing. The text data contains a text file, which is a movie review collected from the movie review site, 'IMDB'.

- Dataset 5 & 6 (large data 1 & 2): In this study, large datasets were built manually for the hands-on testing. The data set 1 (Survey example data), and the data set 3 (Time series) were manually made large by increasing the numbers of rows and columns. The newly made datasets were named as large data 1, and large data 2 respectively. Preprocessing of large data sets belong to big data characteristics volume, as described in Chapter 2.1.1.

The datasets described above includes structured, semi-structured and unstructured data reflecting the variety of data, and dataset 5 & 6 reflects volume of data, not large volume though. The velocity of data was not a considered for the hands-on testing, as the main focus of the thesis was preprocessing.

Table 19 summarizes the datasets:

No.	Dataset name	File type	Number of attributes	File size	Value types
1	Survey example	CSV	11	110KB	Numeric
2	Mobile usage behavior	CSV	66	56KB	Numeric, Nominal, Polynomial, Binominal
3	Time series	CSV	132	551KB	Numeric
4	Text data	TXT	NA	6KB	Text
5	Large data 1	CSV	400	69000KB	Numeric
6	Large data 2	CSV	700	50000KB	Numeric

Table 19: Datasets descriptions for hands-on testing

- Preprocessing task 1 (on Survey example dataset): Preprocessing task 1 was performed on dataset 1. The preprocessing task was to create 39 new attributes, naming 1, 2, 3 up to 39, which will present the count of each response by the participants. For example, the resulting attribute '1' will show, how many '1' were chosen by a participant in the corresponding row, attribute '2' will present the numbers of '2' selected by that participant, and so on. The preprocessing task 1 is categorized as a simple preprocessing task. Preprocessing task for

large data 1 was kept the same. This kind of preprocessing task on this type of data set will allow the operators to find out one user's satisfaction level about their services and products, and the service or product's popularity.

- Preprocessing task 2 (on Mobile usage behavior dataset): This preprocessing task is categorized as advanced level preprocessing, where the tasks were:
  - *To define the data types correctly:* Because of unusual values, tools typically categorize numeric parameters as nominal, polynomial, or string. The first task was to define the data types correctly.
  - *Declare missing values:* The task was to declare the unusual values (n.a., not available) as missing values, so that the tools handle those as missing values other than some random text value.
  - *Impute missing value:* This task involved imputing or replacing all the missing values of a certain attribute with the average value of available values. This is an important preprocessing task, as some actual processes, e.g. clustering cannot read missing values.
  - *Remove unusable variables:* Sometimes some parameters have lots of missing values, which make them useless. In this step the task was to remove all the attributes which have more than 40% missing values.
  - *Outlier detection:* Detecting outliers allows the user to be sure about the quality of the data. In this step the task was to check whether any attribute contains outliers or not.
  - *Filter out missing values:* The task was to filter out rows which contain missing value for one specific attribute.
  - *Aggregation of two parameters:* The task was to aggregate two attributes to generate new one, with their mean values.
  - *Resample the dataset:* The task was to randomly select 50 percent of the dataset by sampling and reduce the data.
  - *Principal Component Analysis (PCA):* The task was to decreasing attributes using PCA, with PCA fixed number as 3.

  This kind of preprocessing task on this type of dataset might help the operators to perform actual analysis such as, correlation analysis, and clustering to know about a subscriber's usage behavior and interest.

- Preprocessing task 3 (on Time series dataset): The task was to aggregate each first three attributes by summing them, and generate a new attribute. In the result, two attributes will be deleted leaving only the first of the three attributes, and the resulted attribute. The data set contains numbers of attributes, so it is time consuming and difficult to aggregate each three attributes, and filter out two manually. If a tool is able to perform this task through automatic iteration or loop, it passes this task otherwise fails. Preprocessing task for large data 2 was kept the same. Preprocessing task like this will allow the operators to know the usage behavior of a subscriber. Insights, such as how much time a subscriber usually spends on specific services, and which timeslot is the peak time for him/her can be generated through this kind of data preprocessing.
- Preprocessing task 4 (on Text data): The task was to preprocess the text file, which contains a) read the text file, b) tokenize it if possible, c) filter out the stop words, and d) build a box of words. Few sentiment analyses were also done with RapidMiner and KNIME with the additional positive and negative review data sets. As this thesis is focusing on only preprocessing of data, the sentiment analyses will not be presented in this thesis. This kind of preprocessing on this type of dataset will let the operators know about the sentiment of a subscriber for a product, and build a learning model to use it for further sentiment analysis on other texts. Text data analysis can also be used in different use cases such as network equipment's error message analysis for network improvement, subscribers' complaint analysis for customer care service improvement.

6.2. *Tools performance evaluation criteria*

Features list for a preprocessing tool is presented in Chapter 5 of this thesis. The tools are selected based on those features. In the hands-on testing result, only performance level criteria will be concentrated. The performance level criteria used in the result are shortly described below.

Complexity: To successfully perform a preprocessing task, every tool requires a series of steps to be selected by the user. Complexity describes the easiness (e.g. easy to learn, time spend and complexity to build the process) to perform the preprocessing tasks.

This criterion has three possible outcomes in a complexity order, named as complex, moderate and less complex.

Exact intendant output: According to the preprocessing tasks, the required outputs were also defined. Sometimes, one tool can perform the preprocessing task but cannot present the output exactly how it was intendant to.

This criterion has to possible outcomes, namely achieved and not achieved.

Memory taken: This criterion describes how much system memory the tool consumes while running the process. Some tools, such as RapidMiner, and Orange in case of large dataset requires so much memory that the system gets slower. Sometimes, some tool requires memory even more than the system memory to run big process on large data, and so the tool fails to perform the process.

This criterion has numeric values which represents the memory taken by the tools in kilobytes.

Time taken to execute: This criterion presents the timing of the tools to execute the processes. The time taken to build the process or load the data was not considered in this criterion. Some tools such as RapidMiner require loading the data separately, and then execute it for some cases e.g. preprocessing task 2. On the other hand, some tools such as KNIME are capable of loading the data and execute the process as a single process. This criterion presents the total timing e.g. loading time executing the process. This is important criterion especially for big data preprocessing.

This criterion has numeric values which represent the time taken by a tool to execute a process in seconds (s) or minutes (m).

User intervention required: This criterion reflects the capability of a tool to perform specific preprocessing tasks more automatically, with less intervention or manual activities of the user.

This criterion has three possible outcomes in order, naming much, moderate and less.

Result: This criterion summarizes whether a tool was capable of performing the preprocessing task or not.

This criterion has two possible outcomes naming, passed, and failed, baring their usual meaning.

Result Table:

Tools	Criteria	Task 1		Task 2	Task 3		Task 4
		Dataset 1	Large data1		Dataset 3	Large data2	
RapidMiner Version: 5	Complexity	Less complex	Moderate	Moderate	Complex	Complex	Less complex
	Exact intendant output	Achieved	Achieved	Achieved	Not achieved	Not achieved	Achieved
	Memory taken	253276K	1144864K	265156K	783684K	2144346K	783100K
	Time taken to execute	0s	7m	1s	1s	27m	0s
	Result	Passed	Passed	Passed	Passed	Passed	Passed
	User intervention required	Much	Much	Moderate	Less	Moderate	Less
KNIME Version: 2.8.0	Complexity	Moderate	Less complex	Moderate	Complex	Less complex	Less complex
	Exact intendant output	Achieved	Achieved	Achieved	Achieved	Achieved	Achieved
	Memory taken	167000K	181321K	176364K	180008K	182500K	204708K
	Time taken to execute	0s	6m	1s	2s	39m	0s
	Result	Passed	Passed	Passed	Passed	Passed	Passed
	User intervention required	Less	Less	Less	Less	Less	Less
Orange Version: 2.7	Complexity	Complex	Complex	Moderate	Complex	Complex	Less complex
	Exact intendant output	Achieved	NA	Achieved	NA	NA	Achieved
	Memory taken	121424K	168562K	125552K	NA	NA	141744K
	Time taken to execute	NA	NA	NA	NA	NA	NA
	Result	Passed	Failed	Passed	Failed	Failed	Passed
	User intervention required	Much	NA	Moderate	NA	NA	Less
IBM SPSS Statistics Version: 21	Complexity	Less complex	Less complex	Moderate	Complex	Complex	NA
	Exact intendant output	Achieved	Achieved	Achieved	NA	NA	NA
	Memory taken	167064K	181664K	170108K	NA	NA	NA
	Time taken to execute	NA	NA	NA	NA	NA	NA
	Result	Passed	Passed	Passed	Failed	Failed	Failed
	User intervention required	Much	Much	Much	NA	NA	NA

Table 20: Hands-on testing result table

6.3. *Tools performance evaluation*

6.3.1. KNIME

KNIME (Konstanz Information Miner) is an open source data analytics, reporting, and integration platform. KNIME integrates various components for machine learning and data mining through its modular data pipelining concept. A graphical user interface allows series of nodes (user interfaces to perform different tasks, also known as operators in RapidMiner, and widgets in Orange) for data preprocessing, for modeling and data analysis and visualization (KNIME Tech, 2013). KNIME can be downloaded on the personal computer and used free of charge. KNIME is implemented in Java, but allows for wrappers calling other code in addition to providing nodes that allow running Java, Python, Perl, and other code fragments (Wikipedia, 2013).

Performance evaluation based hands-on testing

KNIME was able to perform all the preprocessing tasks (preprocessing tasks 1, 2, 3 and 4) in hands-on testing. It is quite easy to learn, user friendly and flexible for every operating systems.

For the preprocessing task 1, it was easy to build the process on KNIME and the output was perfectly matched with the intendant output.

In the preprocessing task 2, the nodes used in the KNIME were easy to search out from the large numbers of nodes. Few sub-tasks, such as removing variables having more than 40% missing value were best performed by KNIME among all selected tools.

Only KNIME was able to give out the exact intendant output for the preprocessing task 3. As it was a big preprocessing task and needed to be performed on large numbers of variables, ‘loop’ was needed to execute. To do this preprocessing task, numbers of nodes were needed to use and the output was perfect.

KNIME was able to perform text analysis (preprocessing task 4) with its ‘Text processing’ extension. The text preprocessing task was completed by KNIME successfully.

For preprocessing on large datasets, KNIME performed the best. The reasons were the following: 1) KNIME actually consumes very less memory while running a big process on large data. At first such large dataset and big preprocessing task was tried as a test process, that only KNIME out of all tools was able to finish it in 72 hours without hampering the performance of the computer. In this test process all the other tools were failed to perform because of memory requirements and/or their less capability 2) KNIME can load large data and perform preprocessing task at a time, separate

data loading is not required. 3) In KNIME robust big data extensions are available for distributed frameworks such as Hadoop.

KNIME was able to cope with the volume and variety of the datasets used for hand-on testing. It also has the capability to perform real-time analytics to cope with velocity of data.

All the processes screenshots are provided in Appendix 6A.

6.3.2. RapidMiner

RapidMiner, formerly known as YALE (Yet Another Learning Environment), is an environment for machine learning, data mining, text mining, predictive analytics, and business analytics (RapidMiner, 2013). It is used for research, education, training, application development, and industrial application. In poll by KDnuggets, RapidMiner ranked second in analytics tools used for real projects in 2009 (KDnuggets, 2009), first in 2010 (KDnuggets, 2010), third in 2012 (KDnuggets, 2012), and again first in 2013 (KDnuggets, 2013). RapidMider is written in Java programming language and is a free of cost tool.

Performance evaluation in hands-on testing

Preprocessing task 1 was successfully completed using RapidMiner. It took several calculative functions to perform manually for this task, but at the end it provided the intendant output.

RapidMiner well performed the preprocessing task 2 and was able to give the exact intendant output, but required moderate user intervention to perform few sub-tasks, such as ‘remove attributes having more than 40% missing values’. It was a large process with numbers of nodes (see Figure 19) to perform all the preprocessing sub-tasks.

Figure 19 is the screen shot of preprocessing task 2 performed on RapidMiner. Small rectangles are the nodes and the lines connect them. The left upper portion of the Figure shows the available nodes under certain domains or extensions, such as Text Processing for user’s use; and the lower portion are the repositories to import and export processes. The right upper portion shows the parameters of selected nodes, e.g. in this case the process is selected. The lower right portion shows the help and comments sections. The process was built by importing different nodes, e.g. ‘Declare Missing Values’, ‘Replace Missing Values’ according to the task requirements.

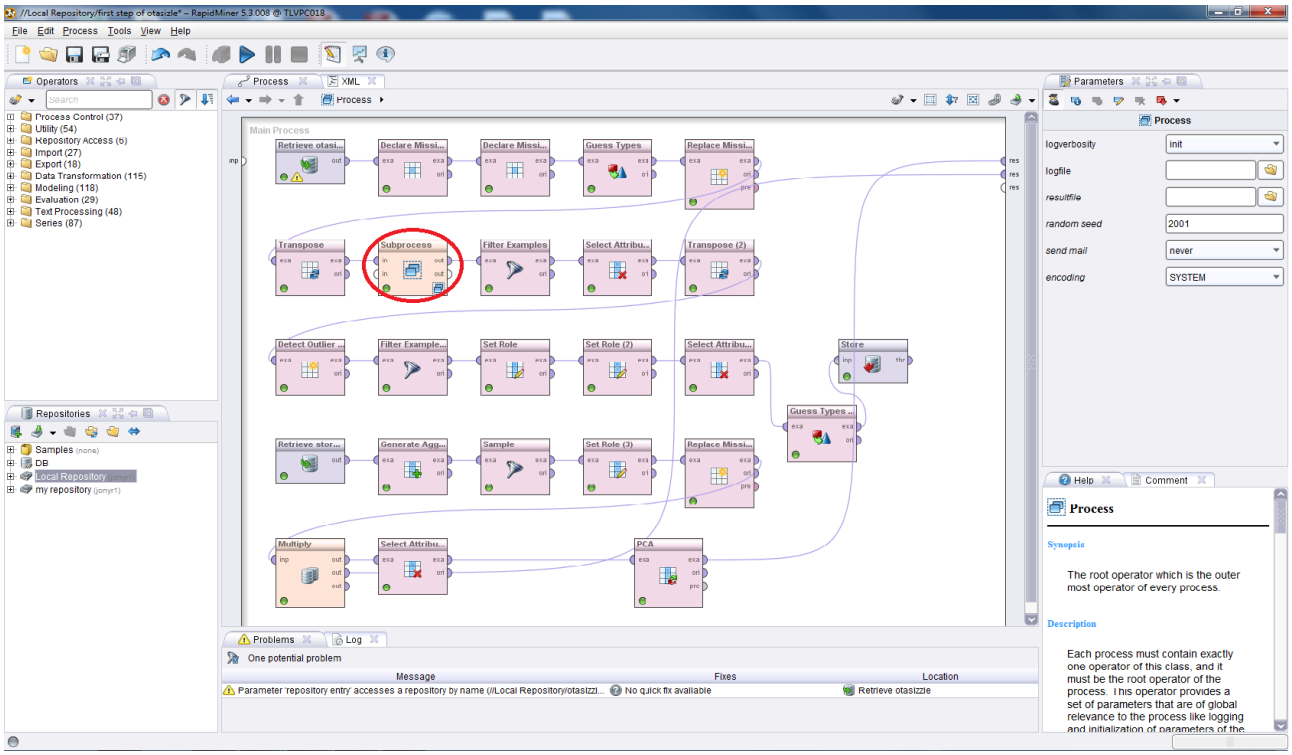

Figure 19: An example of a large process which requires a numbers of nodes

When the process is built and proper connections are done along with one connection to the output, the ‘play’ button executes the process and presents output in the output window. The red marked node is a sub-process, which includes another process under it, the separate sub-process window screen shot has been presented in Appendix 6A.

Preprocessing task 3 was also easy to perform with RapidMiner. But the exact intendant output was not achieved in this case. According to task requirement, the resulting columns were needed to be renamed with the three aggregating columns automatically. Because of discontinuity in the names of attributes, it was difficult to perform it in RapidMiner.

RapidMiner is able to perform text processing with its ‘Text processing’ extension. In fact, RapidMiner has numbers of text processing operators such as extract, tokenize, stem and many more. RapidMiner has most extensive text processing functionality among the tools this study selected for hands-on testing.

For large data preprocessing tasks, RapidMiner was able to perform both the tasks, but it consumes lots of memory of the system. ‘Radoop’ is an extension for RapidMiner, which is built for editing and running ETL, data analytics and machine learning process over Hadoop (Prekopcsak, et al., 2013) for big data.

The hands-on testing results also indicate that RapidMiner is capable of coping up with variety and volume of data, if required memory is available. RapidMiner is also efficient to handle velocity of data.

All the processes screenshots are provided in appendix 6B.

6.3.3. Orange

Orange is a component-based data mining and machine learning software suit, featuring a visual programming front-end for explorative data analysis and visualization, and python bindings and libraries for scripting (Wikipedia, 2013). Orange contains a set of components for data preprocessing, feature and filtering, modeling, and exploration techniques. It is implemented in C++ and Python and is a 100% open source tool.

Performance evaluation in hands-on testing

Preprocessing task 1 was successfully performed by Orange. It required user intervention to specify the conditions manually for each new attributes.

Orange was very efficient to perform the Preprocessing task 2. All the process steps, other than removing variables having more than 40% missing values were automatic. Orange in fact has a very useful widget named ‘Select data’, which counts the percentage of the missing values for every variable and few clicks can remove those attributes which have more than 40% missing values. Along with the attributes, it also automatically deletes the rows which contain such numbers of missing values, which was not intendant for this task. Therefore, the removing process was done manually with the counts collected from the ‘Select data’ node.

Preprocessing task 3 was too complex to successfully perform on Orange. Preprocessing task 3, as an advanced level preprocessing task required automatic iteration. Orange does not have any widget for e.g. iteration or loop.

Orange also has extensions like ‘Text Mining’ and ‘Textable’ for text analysis. The Preprocessing task 4 was completed by Orange using ‘Text Mining’ extension.

Orange typically requires large amount of system memory to execute a process on large dataset. It was unable to perform the preprocessing tasks on large dataset with available system memory of the system used for the hand-on testing. Orange does not have a Hadoop extension yet.

Orange performs the tasks step by step and does not execute the whole process at a time, so the 'time taken to execute' criterion was not applicable.

With capability of text processing, Orange can cope with variety of data. It can also cope with volume of data, if sufficient amount of memory is available. Orange can cope with the velocity of data also.

All the processes screenshots are provided in Appendix 6C.

6.3.4. IBM SPSS Statistics

IBM SPSS Statistics is a commercial software package for statistical analysis. Companion products in the same family are used for survey authoring and deployment, data mining (IBM SPSS Modeler), text analytics (IBM SPSS Text Analytics) (Wikipedia, 2013). IBM SPSS Modeler has a groovy GUI and is good for data analytics. On the other hand, IBM SPSS Statistics is better for statistical analysis which requires much manual activities. As a preprocessing tool, IBM SPSS Statistics is easy to learn and work with, than IBM SPSS Modeler. IBM SPSS Modeler also has extension for IBM SPSS Statistics to import/export process, so it can import and analyze processes which have been performed on IBM SPSS Statistics. IBM SPSS Statistics is built on Java platform.

For preprocessing task 1, IBM SPSS Statistics was the easiest to learn and quickest to build, among the all four chosen tools. It was easy to build the process and the output was exactly what it was intended for.

Preprocessing task 2 was successfully performed by IBM SPSS Statistics, it was easy but a bit time consuming compared to other tools as the process contained many manual steps.

Preprocessing task 3 was complex to solve with IBM SPSS Statistics. IBM SPSS Statistics requires additional Java coding to accomplish the Preprocessing task 3.

SPSS has a separate product for text processing called IBM SPSS Text Analytics.

For preprocessing on large datasets, IBM SPSS Statistics was able to load both the data sets. The large dataset 1 preprocessing task was successfully done with IBM SPSS Statistics, but it consumed a lot time due to the manual steps it required. The large data set 2 preprocessing task was too complex to solve with IBM SPSS Statistics, the reason was the same as preprocessing task 3.

IBM SPSS Statistics performs the tasks step by step, so the 'time taken to execute' criterion was not applicable for it.

IBM Infosphere is a big data platform based on Hadoop.

IBM SPSS Statistics can well suit for the volume and velocity of data. The Hands-on testing shows, how efficiently it can handle large amount of data. It can handle unstructured data as well, but for text preprocessing IBM SPSS Text Analytics would be a better alternative.

Processes screen shots are presented in Appendix 6D.

The four tools' performances were analyzed and evaluated by hands-on testing. Table 20 shows the differences among the tools by presenting the capability of each tool to perform the preprocessing tasks efficiently. Finally, it can be concluded that KNIME and RapidMiner performed best on the hands-on testing, where Orange and IBM SPSS Statics also survived most of the tasks.

7. Conclusion

This section is divided into four parts. The first part summarizes the results of the thesis. The second part is the assessment of the results and the third part is a discussion on the exploitation of the results. The section is concluded by addressing potential future researches.

7.1. *Results*

This research has been able to provide results in the form of two findings and answers to the research questions. Firstly, this thesis identified potential big data use cases and corresponding functional requirements for telecom industry. Secondly, it presented the list of available data preprocessing tools giving heed to their main focus and rankings and also distinguished two most promising tools for big data preprocessing i.e. KNIME and RapidMiner.

7.2. *Assessment of Results*

This study identifies currently implemented and potential future big data use cases for telecom industry, hence answers the first research question. However, newer use cases are emerging as operators are more focusing on their big data strategy. For example, potential use cases based on location data are emerging currently. As a result, location based services can be introduced as a separate use case domain. Additional use cases will increase the reliability of the use case table. Furthermore, the study also presents corresponding feature requirements and typical data types required for the use cases. However, practical implementation of the use cases might require some additional data. The use cases might also vary in different countries for different operators based on some criteria, such as numbers of subscribers, available services and technical specifications.

The tools-list presented in Table 17 provides the name, popularity rankings, personal qualitative rankings and main focuses of the tools. The tools-list was minimized to 20 selected tools based on certain criteria described in Chapter 5.2. The tools, which were not selected, might also have some rich functionality and preprocessing capability, which were not visible based on those criteria. The selected tools were scored based on their preprocessing functionalities by preliminary testing. In few tools, finding proper node for preliminary testing was difficult, which might has provided some misjudged scoring results. The personal qualitative rankings might also vary in some cases, as features e.g. ‘easy to learn’ differ from user to user.

The hands-on testing in this thesis were sort of simulation of big data preprocessing. The datasets used for hands-on testing reflect typical data used in telecom industry. The large datasets (dataset 5

& 6) represent the volume characteristic of big data, but not high-volume. This research has only managed to perform hands-on testing on datasets that are some tens of megabytes in size.

Few tools, such as RapidMiner typically consumes large amount of system memory to execute complex processes on large datasets with default in-memory (RAM-centric) analytics engine, which also restricts the maximum amount of accessible data points to 25,000,000 (RapidMiner, 2013). However, RapidMiner is also capable of using other types of analytical engines, such as in-database engine (enterprise edition only) and in-Hadoop engine, which are capable to handle more data points e.g. around 50,000,000. KNIME is capable of performing simple data preprocessing tasks on around 176 million rows with a 4 cores 8 GB RAM system (Silipo & Winters, 2013). IBM SPSS Statistics can hold up to 2 billion columns or rows in a dataset, depending on the system capacity (IBM, 2013). Orange also loads all data in the memory by default, therefore consumes large amount of system memory. All the four tools are expected to handle e.g. 10 GB datasets with 64-bit system consisting of 4 cores and 16 GB RAM.

Datasets used for the hands-on testing all together well represent variety of data. Real-time analytics (which represents the velocity characteristic of big data) and its importance were described in detail in this thesis. However, it was not included in the hands-on testing. Real-time analytics requires building a process model on a tool. As a consequence, the tool can be used as real-time analytics engine, where data directly come from the data sources, such as sensors and get processed in real-time. As the main focus of the thesis was preprocessing of data, no such model was built for hands-on testing.

The preprocessing tasks used for hands-on testing involved several sub-tasks and well fit for functionality and usability measurement of a data preprocessing tool. Therefore, the hands-on testing results answer the second research question of this thesis.

KNIME, RapidMiner and Orange are open source tools and allow the users to create new nodes according to their requirements to extend the functionality. IBM SPSS Statistics also provides option to further modify the analyses by using command syntax. However, IBM SPSS Statistics is a commercial tool. Therefore, it might not be a cost-efficient solution for small organizations. Other commercial tools also have rich functionality of data preprocessing but not been tested because of unavailability. The selected tools can also be used in other industries than telecom industry for data preprocessing solutions.

7.3. *Exploitation of Results*

The results of this study can be exploited to four different areas as follows:

1. The use case table offers telecom operators a cookbook while planning business strategies where big data is concerned.

2. Table 17 presents a comparison ranking among available data analysis tools according to their focused area, which might be useful for researchers and industry experts to choose data analysis tools according to their application requirements.

3. The comparison scoring of available preprocessing tools based on their preprocessing capability will help the researchers or industry experts to choose a preprocessing tool according to their requirements. For example, an expert who usually works on datasets with many missing values might choose a tool which got a good score in missing value analysis e.g. RapidMiner.

4. The hands-on testing gives a relevant performance results for parties, who are selecting a data preprocessing tool for their project.

7.4. *Future Research*

This thesis offers potential research scopes for the researchers and industry experts by highlighting industrial big data usage and preprocessing solutions.

As the field of big data is still in a great state of change and development, the possibilities for research in this area are extensive and the interest on the findings will certainly rise if the adoption of big data rises as expected. The big data use cases list for telecom industry (Table 22, Appendix 3A) might provide potential future research scopes. Most of the listed use cases are not implemented yet and thus give room for future research works, where each use case can be potential research topic. The use case table can also be optimized as a further research.

Preprocessing of data has been a broad research topic for long time. New tools, new technologies and newer types of data are making it more challenging and broader. There are potential scopes of future work on data preprocessing. Tools for data preprocessing are also emerging currently to meet the newer requirements. This thesis has managed to provide hands-on testing results considering two (volume and variety) characteristics of big data. Preprocessing of big data including all three characteristics (volume, variety and velocity) will be a potential future research work.

Commercial tools are also getting popular with their rich functionalities and capabilities. However, this thesis has managed to test three available commercial tools. Actual hands-on testing and comparison of the other commercial tools is expected to be a potential research topic. New open source and commercial tools, with new features and new capabilities are emerging and may replace the current tools. Capability comparison among future and current tools will also be a potential topic for research work. Practical application of the selected preprocessing tools is another potential future research scope of this thesis.

8. Bibliography

Acker, O., Blockus, A. & Pötscher, F., 2013. *Benefiting from Big Data: A New Approach for the Telecom Industry,* s.l.: Booz & Company.

Ackloff, R., 1989. From Data to Wisdom. *J. Applied Systems Analysis,* Volume 16, pp. 3-9.

actian, 2012. *Hadoop's Limitations for Big Data Analytics,* s.l.: ParAccel.

Aginsky, A., 2012. Is big data the next big thing for telecom?. *EMEA,* pp. 34-35.

Anon., 2013. *ETL.* [Online]
Available at: http://www.dataintegration.info/etl

Apache, 2012. *Hadoop.* [Online]
Available at: http://hadoop.apache.org/
[Accessed July 2013].

AT Kearney, 2013. *Big Data and the Creative Destruction of Today's Busniness Model,* s.l.: s.n.

Banerjee, A., 2011. Addressing 'Big Data' Telecom Requirements for Real-Time Analytics. *Heavy Reading,* March.

Bellaachia, A., n.d. *Data Preprocessing.* s.l.:s.n.

Bellinger, G., Castro, D. & Mills, A., 2004. Data, Information, Knowledge, and Wisdom.

Bernstein, J. H., 2011. The Data-Information-Knowledge-Wisdom Hierarchy and its Antithesis. *North American Symposium on Knowledge Organization,* pp. 68-75.

Beyer, M. A. & Laney, D., 2012. *The Importance of 'Big Data': A Definition,* s.l.: Gartner.

Biehn, N., 2013. *The Missing V's in Big Data: Viability and Value.* [Online]
Available at: http://www.wired.com/insights/2013/05/the-missing-vs-in-big-data-viability-and-value/

Botteri, P., 2012. *Eastern European Champions & the 4 V's of Big Data.* [Online]
Available at: http://cracking-the-code.blogspot.fi/2012/10/eastern-european-champions-4-vs-of-big.html

Boyd, D. & Crawford, K., 2011. Six Provocations for Big Data. *Social Sceince Research Network,* 13 September.

Brynjolfsson, E., Hitt, L. M. & Heekyung Hellen, K., 2011. Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?. *Social Science Research Network,* 22 April.

Chen, M. et al., 2008. Data, Information, and Knowledge. *Computer Graphics and Applications, IEEE,* 29(1), pp. 12-19.

Cisco, 2013. *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012–2017,* s.l.: s.n.

CITO Research, 2012. *Big Data Insights in a Rush,* s.l.: s.n.

Demchenko, Y. et al., 2012. *Addressing Big Data Challenges for Scientific Data Infrastructure.* Taipei, Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on.

Doctorow, C., 2008. Big Data: Welcome to the petacentre. *Big Data: Science in the petabytes era,* pp. 16-21.

Egham, 2013. *Gartner Says Worldwide Mobile Phone Sales Declined 1.7 Percent in 2012,* London: Gartner.

Elisa, 2012. *Elisa Annual Report 2012,* s.l.: Elisa Oyj.

Elisa, 2013. *Elisa Vahti.* [Online]
Available at: http://www.elisa.fi/vahti/
[Accessed September 2013].

Ericsson, 2012. *Traffic and Market Report,* s.l.: s.n.

Ericsson, 2013. *Big Data Analytics,* s.l.: Ericsson.

Famili, F., Shen, W.-M., Weber, R. & Simoudis, E., 1997. Data Pre-Processing and Intelligent Data Analysis. *International Journal on Intelligent Data Analysis,* I(1).

Feinleib, D., 2012. The Big Data Landscape. *Forbes,* 19 June.

Friedman, V., 2008. Data Visualization and Infographics. *Smashing Magazine,* 14 January.

Gartner, 2012. *Gartner Reveals Top Predictions for IT Organizations and Users for 2013 and Beyond,* Orlando: Gartner.

Han, J., Kamber, M. & Pei, J., 2011. Data Preprocessing. In: *Data Mining: Concepts and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems).* 3rd ed. s.l.:s.n., pp. 83-123.

Hunter, P., 2013. Journey to the centre of Big Data. *Engineering and Technology, IEEE Magazine,* 27 April, pp. 56-59.

IBM, 2010. *Globe Telecom: Gaining marketing agility with smart promotions,* s.l.: s.n.

IBM, 2012. *Analytics: The real-world use of Big Data,* s.l.: s.n.

IBM, 2012. *Tier 1 Mobile Service Provider transforms call centers with real timeaccess to customer and product data,* s.l.: IBM Corporation.

IBM, 2013. *What is the Maximum Number of Columns / Variables in Statistics?.* [Online]
Available at: http://www-01.ibm.com/support/docview.wss?uid=swg21479527
[Accessed November 2013].

Infosys, 2013. *Big Data: Challenges and Opportunities,* s.l.: s.n.

ITU, 2013. *TELCO BIG DATA: 2013 WILL BE A STRATEGIC YEAR FOR MOBILE OPERATOR,* s.l.: International Telecommunication Union.

Ji, C. et al., 2012. *Big Data Processing in Cloud Computing Environments.* San Marcos, Pervasive Systems, Algorithms and Networks (ISPAN), 2012 12th International Symposium on.

Jony, R. I., 2013. *Experts' Interviews.* Espoo, Helsinki: s.n.

Kaisler, S., Armour, F., Espinosa, J. A. & Money, W., 2013. *Big Data: Issues and Challenges Moving Forward.* Wailea, Maui, HI, s.n., pp. 995 - 1004.

Kalakota, R., 2012. *Sizing "Mobile + Social" Big Data Stats.* [Online]
Available at: http://practicalanalytics.wordpress.com/2012/10/22/sizing-mobile-social-big-data-stats/

Karikoski, J., 2012. Handset-Based Data Collection Process and Participant Attitudes. *International Journal of Handheld Computing Research (IJHCR),* III(4), pp. 1-21.

KDnuggets, 2009. *Data Mining Tools Used Poll.* [Online]
Available at: http://www.kdnuggets.com/polls/2009/data-mining-tools-used.htm
[Accessed September 2013].

KDnuggets, 2010. *Analytic Tools Used Poll (May 2010).* [Online]
Available at: http://www.kdnuggets.com/polls/2010/data-mining-analytics-tools.html
[Accessed October 2013].

KDnuggets, 2012. *Analytics, Data mining, Big Data software used (May 2012).* [Online]
Available at: http://www.kdnuggets.com/polls/2012/analytics-data-mining-big-data-software.html
[Accessed 2013].

KDnuggets, 2012. *Data Mining application in 2012.* [Online]
Available at: http://www.kdnuggets.com/polls/2012/where-applied-analytics-data-mining.html

KDnuggets, 2013. *Analytics/Data mining software used?.* [Online]
Available at: http://www.kdnuggets.com/polls/2013/analytics-big-data-mining-data-science-software.html
[Accessed 22 September 2013].

Kekolahti, P., n.d. s.l.:Unpublished work.

Kelly, P. M. & White, J. M., 1993. *Preprocessing remotely sensed data for efficient analysis and classification.* s.l., Applications of Artificial Intelligence 1993: Knowledge-Based Systems in Aerospace and Industry.

KNIME Tech, 2013. *KNIME.* [Online]
Available at: http://www.knime.org/
[Accessed September 2013].

Lande, J. v. d., 2013. *Big Data Analytics: How CSPs can can generate profits from thier data,* s.l.: Analysys Mason.

Madden, S., 2012. From Databases to Big Data. *Internet Computing, IEEE,* 16(3), pp. 4 - 6.

Madsen, L., Meggelen, J. V. & Bryant, R., n.d. Call Detail Records. In: *Asterisk: The Definitive Guide.* 3rd ed. s.l.:s.n.

Malik, O., 2011. Internet of things will have 24 billion devices by 2020. *GIGAOM*, 13 October.

McGuire, T., 2013. *Making data analytics work: Three key challenges* [Interview] (21 March 2013).

McKendrick, J., 2013. *2013 BIG DATA OPPORTUNITIES SURVEY,* s.l.: Unisphere Research.

McKinsey & Company, 2011. *Big Data: The next frontier for innovation, competition, and productivity,* s.l.: s.n.

Metascale, 2013. *Why Hadoop.* [Online]
Available at: http://www.metascale.com/why-metascale/why-hadoop
[Accessed August 2013].

Middleton, J., 2013. *Turkcell using Big Data to track faults.* [Online]
Available at: http://www.telecoms.com/173682/turkcell-using-big-data-to-track-faults/
[Accessed 23 September 2013].

Miller, H. & Mork, P., 2013. From Data to Decisions: A Value Chain for Big Data. *IT Professionals, IEEE Computer Society*, 4 February, pp. 57 - 59.

Moorman, C., 2013. The Utilization Gap: Big Data's Biggest Challenge. *Forbes,* 17 March.

Orange (Software), 2012. *File Widget Files size limit.* [Online]
Available at: http://orange.biolab.si/forum/viewtopic.php?f=4&t=1580
[Accessed November 2013].

Orange (telecommunications), 2013. *Joining up healthcare,* s.l.: Orange Healthcare.

PC Magazine, 2013. *Definition of: Big Data.* [Online]
Available at: http://www.pcmag.com/encyclopedia/term/62849/big-data

Porter, M. E., 1985. *Competitive Advantage: Creating and Sustaining Superior Performance.* s.l.:Simon and Schuster.

Pradhan, M., 2012. *What lies at the core of Hadoop?.* [Online]
Available at: http://blog.enablecloud.com/2012/06/what-lies-at-core-of-hadoop.html
[Accessed September 2013].

Prekopcsak, Z., Makrai, G., Henk, T. & Gaspar-Papanek, C., 2013. *Radoop: Analyzing Big Data with.* Budapest, RapidMiner Community Meeting and Conference.

RapidMiner, 2013. *RapidMiner.* [Online]
Available at: http://rapidminer.com/
[Accessed September 2013].

RapidMiner, 2013. *RapidMiner & Big Data – How Big is Big?.* [Online]
Available at: http://rapidminer.com/2013/07/31/rapidminer-big-data-how-big-is-big/
[Accessed 26 November 2013].

Sagiroglu, S. & Sinanc, D., 2013. *Big Data: A review.* s.l., s.n., pp. 42-47.

Schmidt, E., 2010. *Every 2 Days We Create As Much Information As We Did Up To 2003,* s.l.: TechCrunch.

Sicular, S., 2013. Gartner's Big Data definition consists of three parts. *Forbes,* 27 March.

Silipo, R. & Winters, P., 2013. *Big Data, Smart Energy, and Predictive Analytics-Time Series Prediction of Smart Energy Data ,* s.l.: KNIME.

Simitsis, A., 2003. Modeling and managing ETL processes.. *VLDB PhD Workshop.*

Smith, M., Szongott, C., Henne, B. & Voigt, G. v., 2012. *Big data privacy issues in public social media.* Campione d'Italia, Digital Ecosystems Technologies (DEST), 2012 6th IEEE International Conference on.

Soubra, D., 2012. *The 3Vs that define Big Data.* [Online]
Available at: http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data?id=6448529%3ATopic%3A20334&page=2

Statsoft, 2013. *Rexer Analytics.* [Online]
Available at: http://www.statsoft.com/Company/About-Us/Reviews/2013-Published-Reviews#rexerhighlights2013/
[Accessed 2013].

Stein, A., 2012. *Big Data and Analytics, The Analytics Value Chain – Part 3,* s.l.: SteinVox.

Tanasa, D. & Trousse, B., 2004. Data Preprocessing for WUM. *Advanced Data Preprocessing for Intersites Web, published in IEEE Intelligent Systems,* II(19), pp. 59-65.

TeliaSonera, 2013. *Spotify music on the computer, mobile, TV and almost anywhere!.* [Online]
Available at: http://www.teliasonera.com/en/innovation/entertainment/2011/4/draft/
[Accessed November 2013].

The Economist, 2010. *Data, data everywhere,* s.l.: s.n.

Tstat, n.d. [Online]
Available at: http://tstat.tlc.polito.it/measure.shtml#log_mm_complete

Wallin, S. & Landen, L., 2008. *Telecom Alarm Prioritization.* Okinawa, Advanced Information Networking and Applications - Workshops, 2008. AINAW 2008. 22nd International Conference on.

Weathington, J., 2012. Big Data Defined. *TechRepublic*, 3 September.

Weiss, G. M., 2005. Data Mining in Telecommunications. In: *Data Mining and Knowledge Discovery Handbook.* s.l.:s.n., pp. 1189-1201.

Wikipedia, 2013. *Apache Hadoop.* [Online]
Available at: http://en.wikipedia.org/wiki/Apache_Hadoop
[Accessed August 2013].

Wikipedia, 2013. *Big Data.* [Online]
Available at: http://en.wikipedia.org/wiki/Big_data

Wikipedia, 2013. *KNIME.* [Online]
Available at: http://en.wikipedia.org/wiki/KNIME
[Accessed September 2013].

Wikipedia, 2013. *Orange (software).* [Online]
Available at: http://en.wikipedia.org/wiki/Orange_(software)
[Accessed September 2013].

Wikipedia, 2013. *SPSS.* [Online]
Available at: http://en.wikipedia.org/wiki/SPSS
[Accessed September 2013].

Wikispaces, 2013. *Weka.* [Online]
Available at: http://weka.wikispaces.com/Can+I+use+CSV+files%3F
[Accessed September 2013].

Wu, X., Zhu, X., Wu, G.-Q. & Ding, W., 2013. Data Mining with Big Data. *Knowledge and Data Engineering, IEEE Transactions on,* pp. 1-25.

Appendices

Appendix 1A

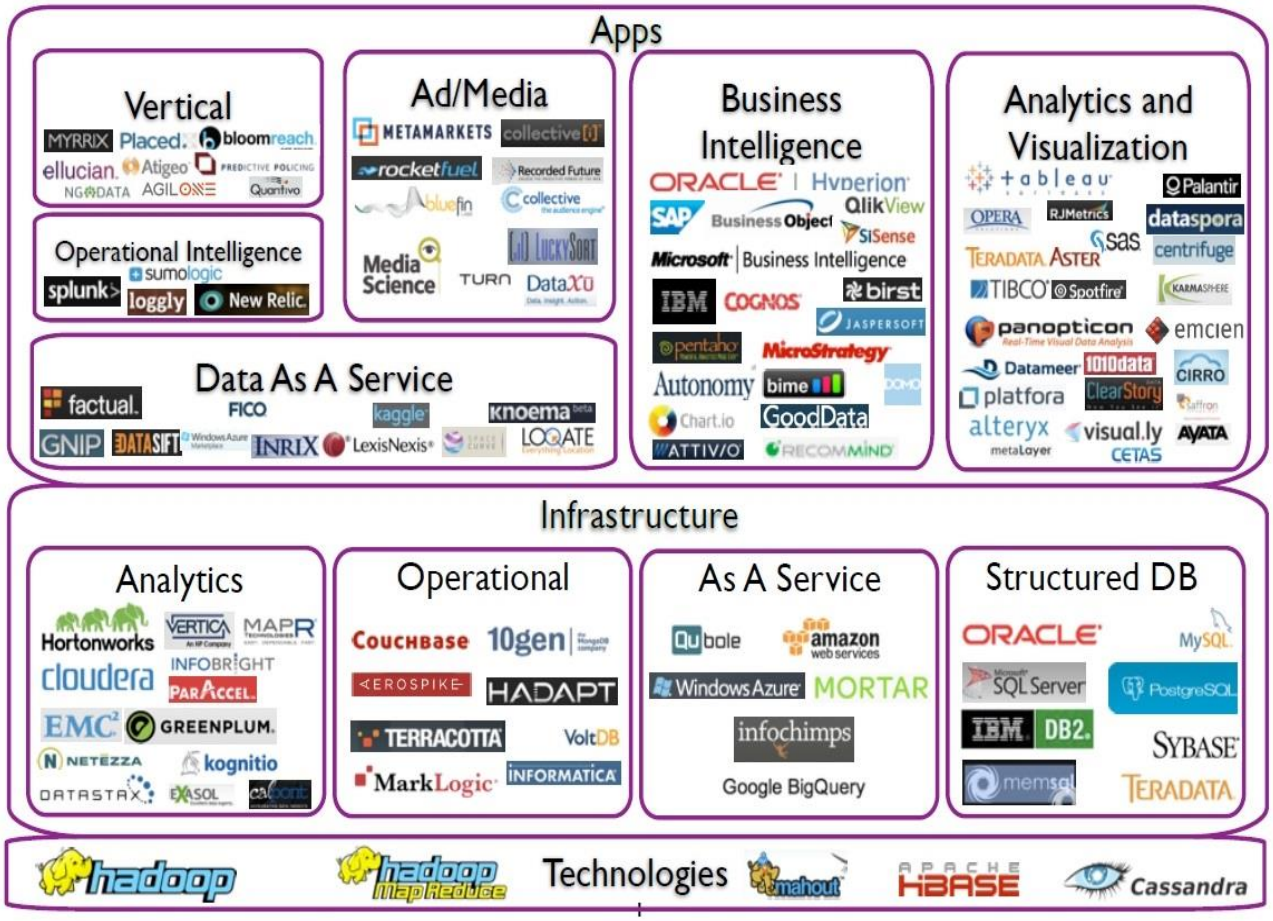

Figure 20: Big data landscape **(**Feinleib**,** 2012**)**

Appendix 2A

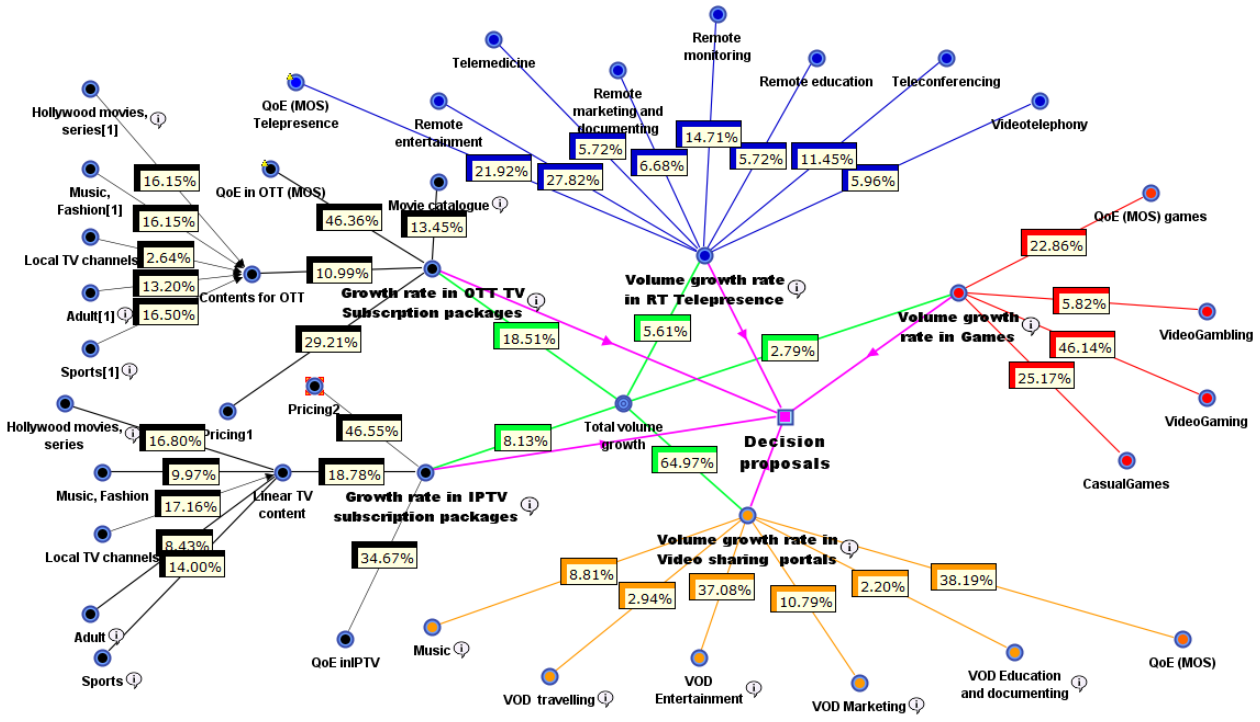

Figure 21: Traffic volume from some sources for end user **(**Kekolahti**,** n.d.)

Appendix 2B

Banking	Transportation	Governmental	Retail
Transactions	Map	Tax payment	Product info
support	Route planning	Online permitting	Price comparison
content service access	Journey planning	Forms and application	Floor maps
video interaction with	Weather forecast	Web mapping	Product research
test drive demos	Location tracking	Paying tickets	Ads
SMS banking	Online booking and	Contact info access	Reading reviews
Bill payment	Pricing	Election info	Photo sharing
QoE	Traffic condition	Municipal courts and	Purchase
	QoE	Licensing	QoE
		Tourism info	
		Public health info	
		Public safety info	
		Parking and community	
		Education	
		QoE	

Table 21: Additional traffic volume sources for end user perspective

Appendix 3A

Domain	Use case name	Short description	Typical data and data sources	Typical requirements
Improvement different kind of characteristics of the network	Improve coverage and QoS	Improve coverage and the quality of service by analyzing the network data while and where required.	Network data (throughput, capacity, coverage), Subscriber data (number of subscriber), CDRs and Location data from network equipment (BTS, VLR, HLR, MSC, BSC etc.)	Capability of collecting data from diverse sources, big data storage capability to store all these data, capability of real time data analysis to learn the current coverage and QoS situation, capability of real time query and access to the data bases to find out how to improve it and capability of real time dash boarding to visualize.
	Network optimization and planning	Plan the network using datasets including previous, present network structure and Optimize the network when and where needed.	Network Planning data (Heights of the base stations, frequencies in use, average radius of the BTS, throughput, capacity , traffic intensity , traffic load, link budget, area data), Network optimization data (antenna specifications like tilt and azimuth, required throughput)	Capability of storing large amount of data, capability of doing predictive analysis to learn where the number of subscriber could grow and how much coverage would be required to meet growing traffic.
	Build smarter network to drive consistent network, high QoE	Building self-optimizing or intelligent network. few operators are working on it at some extent	Network elements interface data, CDRs, probes data, location data	Capability of storing big data, capability of retrospective analysis, capability of real time analysis, capability predictive analysis
	Efficient root cause analysis and Fault isolation	Through proper data analysis it is possible to find the root cause of any event failure and through proper predictive analysis it is possible to isolate the faults even before happening	Operational data from network equipment, data collected from probes, historical data of event failure including the magnitude, location, timing and nature of the faults	Capability of storing big amount of event historical data, capability aggregating diverse data gather all the data required for the analysis, capability of root cause analysis, capability of predictive analysis to learn when and where this kind of failure might happen, capability of real time analysis to prevent the fault to happen
	Acquire next best devices/products for network	By Analyzing the datasets of performance measurement and requirements operators can acquire next best products for their network.	Device data, device log, Market research data, Network requirements data	Capability of in memory historical data analysis to find out which devices are performing poor for a time being, Predictive analysis to learn which would be the new requirements to meet and capability of regression analysis to compare with the available other products or devices

	Network monitoring	Big data analysis can let the operators monitor the whole network. Network monitoring will let the operators improve their decision making and improve the network so on	Network data from probes and interfaces, operational data, CDRs	Capability of collecting data form diverse sources, capability of data aggregation to gather required data, real time data analysis and capability of real time reporting
	Prevent over utilization of network resources	Analyzing the proper data can give an idea on requirements of the resources in the network and where and when the resources are being over utilizing	Network resource data, traffic data, network configuration data, data from probes	capability of data aggregation to aggregate the data form diverse sources, capability of diverse data collection, capability of real time data analysis to monitor real time resource management
	Less power consumption	By proper data analysis Power consumption by the network elements can be reduced	Equipment manual data, operational data, power consumption data	capability of data aggregation, capability of predictive analysis, capability of real time data analysis, capability of real time decision making
	Improve network by deploying mobile cells when needed	In case of concert or some other big public gathering, operators can predict the traffic load by proper analysis of location data and traffic data then can deploy mobile cells temporarily to improve the network to handle the load for certain time	Location data, traffic data, network coverage and throughput data	capability of data aggregation, capability of diverse data collection, capability of proper data correlating and capability of predictive analysis
	Capacity management	Big data analytics can provide operators good hold on network capacity management. Proper data analysis will let the operator know about current capacity, predict future capacity requirement and make decisions accordingly	Traffic data, network elements data, capacity utilization data, data from probes	capability big data storage, data aggregation, real time analysis, real time dash boarding
	Real time visibility in operations	Big data analyzing capability will give the operators the chance to make the live operations visible to them. Real time visibility in operation surely can help the operators improve the network and earn customer satisfaction	Network element data, probes data, CDRs, location data, network configuration data	capability of storing big data, capability of real time data aggregation, capability of real time dash boarding
	Identify network load and bottleneck	How much load the network is handling and what is the limitation can be used to find out the bottleneck	Network traffic data, probes data, capacity data	capability of Real time data analysis, descriptive analysis, capability of data aggregation and capability of diverse data collection

	Prioritizing the alarms automatically	From previous historical data, alarms can be prioritized according to very urgent to no deadline. This will allow the operators to fix or concentrate on the higher prioritized events first when several events or failure occurs simultaneously	Alarm data with severity and associated trouble ticket with manually set priority, redundancy, topology or SLA data, dynamic information available at run-time from RNC, RBS, probes	Capability of big data storage, capability of accessing data bases very quickly, capability of data aggregation to aggregate the event historical data and real time operational data, capability of real time data analysis, capability of finding correlation, capability of learning models and capability of predictive analysis, capability of pattern recognition, capability of text processing
	Network process management	If any parameter of network needs to be changed, operators can do it and measure the performance in real time so that if it is not good for the network it will not affect the network much and can be changed again	Network configuration data, network historical data, network elements data from the probes	Capability of storing big data, capability of real time data aggregation, capability of real time data analysis
	Making work flow more efficient	The work flow of the network can be made more efficient through proper data and information analysis	Network elements data, network configuration data, operational historical data	capability of diverse data collection, data aggregation, in memory analysis, Real time data analysis,
	Network intelligence generation	By reviewing past network event data, network providers can identify the geographic areas, times of day, and specific days of the week, month or year when network performance is poor and improve it	Network configuration data, network historical data, location data	Capability of storing big data, capability of retrospective analysis, capability of real time analysis
	Network performance management	Big data analysis will let the operators to keep a good eye on network performance. Current network performance can be visualize in real time, required network performance can be predicted through popper data analysis	Network configuration data, network performance data, network elements data, data from the probes	Capability of big data storage, real time predictive analysis
	Improve operation management	Operators can use their big data analysis to ensure proper operation management	Service data, operational data from OSS	Diverse data collection, data aggregation, big data storage, predictive analysis, real time dash boarding
	Predictive maintenance	Predictive maintenance (PdM) techniques help determine the condition of in-service equipment in order to predict when maintenance should be performed. This approach offers cost savings, time savings and reduce operation failure as though big data analysis predictive maintenance predicts if there will be any maintenance requirement beforehand	Network elements data, probes data, operation data, history data form databases	Capability of big data storage, capability of data aggregation, capability of real time predictive analysis

	Traffic management	Big data analysis can be used for proper traffic management for the operators. By analyzing proper data, operators can operate their data traffic, call traffic management in real time.	CDRs, mobile positioning data, traffic data	Capability of real time data analysis, descriptive analysis, capability of data aggregation
	Radio reseource management	Radio resource management is one very important task to maintain by the operators for proper service provision. Big data analysis can make radio resource management more automatic and effective	Radio resource data, Network data, Traffic data etc.	Capability of correlation analysis, real time analysis, predictive analysis etc.
Marketing	Provide user specific product and service	Using proper user behavior data analysis one marketing plan can be introduced as provide user specific products and services	subscriber usage profile, billing data, past responses to offers, user behavior data, location data, sentiment data from user social network usage and blogs or forums, from HLR, VLR, user equipment panels, probes etc.	Capability of collecting data from diverse sources, capability of aggregating these data, capability of regression analysis, capability of text processing and capability of sentiment analysis
	Detect business threads	What are the threads to the business can be find out through proper business data analysis	Market survey data, Sales data from the storage database	Capability of subscriber sentiment analysis, capability of predictive analysis to detect the future thread, capability of
	Refining data allowances and pricing	analyzing the user usage data, data allowances and pricing can be refined to earn user satisfaction	User usage data, user behavior data, location data, user social network data, traffic data from HLR, VLR, user equipment panels, probes etc.	Capability of collecting data from diverse sources, capability of aggregating these data, capability of regression analysis, capability of text processing and capability of sentiment analysis
	Geo marketing	Geo marketing is a discipline within marketing analysis which uses geo-location (geographic information) in the process of planning and implementation of marketing activities. According to the location of the subscriber operators can send marketing texts suitable for that specific location	Location data from GPS or device panel, customer data from HLR, reliable consumer profiling data, historical data, service or product specification data	Capability of data aggregation, capability of clustering for subscriber profiling, capability of predictive analysis
	Increase productivity	Productivity can be increased by big data analysis. The reasons can be analyzed in case of less productivity, outcome can be predicted and then optimized accordingly to increase productivity	Sales data, production data, market research data	Capability of storing big data, capability of predictive analysis

	Proactive contract renewals encourage prepaid customer to recharge	When a user´s contract is about to finish, proper service or discount can be offered according to user´s interest and behavior to renew the contract. If subscribers balance is running low and the subscriber is near a recharge location then operators can offer some discount or other offer to encourage him to recharge	User's usage data from handset data, user's behavior data from social networking, package specifications, CDRs for the specific user, SDRs, billing data, balance data	Capability of storing big data, real time decision making capability, capability to access various data sets, capability of data aggregation
	Real time advertisement	According to user usage and current interest operators can provide real time advertisements to the user	Location data from handset data, customer profile data from subscriber data table, subscriber behavior data from social networking or HTTPS, market research data	capability of Real time data processing and decision making to provide the proper advertisement according to the subscriber interest, capability of real time data collection to collect those data in real time, capability of descriptive analysis, predictive analysis
	Intelligent marketing campaign	Big data analytics enable CSPs to better understand their customers and develop subscriber profiles that can be used to create more intelligent marketing campaigns	Marketing history data, location data from handset data, customer data, service data	Capability of correlating various data, capability of collecting data from diverse sources, capability of predictive analysis and decision making to predict proper marketing idea
	Finding shortest path to reach the customers	To reach the customers, by analyzing data find out the most influential customers and use them to reach more customers	Subscriber profiling data, location data from handset data, customer behavior data from social networking site, subscriber usage data from handset data	Capability of storing all subscribers' data, capability of clustering to grouping the subscribers, capability of sentiment analysis and social networking data analysis and text analysis to find out most influential subscriber.
	Check Ad effectiveness	When a new advertising campaign is released into the marketplace, if companies are able to quickly evaluate qualitative consumer responses in real time to determine how the campaign is resonating and with whom, they can quickly tune their campaigns to maximize return on investment	User usage profile, location data, social networking data, market research data, consumer response data	Capability of social networking data analysis and insights generations, capability of text data analysis, capability of sentiment analysis, capability of data aggregation, capability of real time data analysis
	Improve the result of marketing promotions	CSPs can use big data analysis to analyses the past responses to offers to create targeted promotions that customers are more likely to accept	Subscribers usage profile, billing data, past responses to offers	Capability of correlating various data, capability of diverse data collection and aggregation, capability of predictive analysis
	Acquire next best sales and service actions	Through proper data analysis operators can acquire next best sales and service actions like offering new services, an upgrade for an existing service or a service call to address a specific issue	Recent customer interactions across all sales and service such as service calls, SDRs, user usage data and interest data from social networking or handset data	Capability of data aggregation, capability of insight generation, capability of unstructured data analysis, capability of real time analysis, capability of storing big data

	Roaming marketing	Operators can easily identify if one subscriber is in roaming by the location data. By analyzing roaming subscriber activity operators can implement roaming marketing offering special services or information regarding billing, tourism etc.	Subscriber location ate, SDRs, subscriber usage data from the handset panels and HTTPs, subscriber behavior and sentiment data from social networking sites, blogs etc.	Capability of storing big data, capability of predictive analysis, capability of real time data analysis, capability of data aggregation, capability of unstructured data analysis
Security	Fraud management and prevention	Big data analytics allow telecommunication companies to detect and analyze high-velocity fraud activities in real time and take immediate action	Location data from the handset data, subscriber profile data, Data from HLR, VLR, MSC, BSC to detect the SIM, data from CDR and SDR such as basic information of the customers, the number of the calls in certain period, the duration of the call in the certain period, billing data	Capability of data aggregation in real time, capability of diverse data collection, capability of predictive analysis to predict the frequent, capability of learning models
	Real time cyber security monitoring	Operators can provide services like real time cyber security monitoring and prevent the system from attacks	Cyber security data	Capability of storing big data, real time analysis etc.
	Information security	Big data analytics can give CSPs the idea of what kind of attacks are frequent in their information and allow them to predict and prevent those attacks to secure their information	Historical data, network performance data, network behavior data, location data	Capability of real time data analysis, predictive analysis, real time dash boarding etc.
	Prevent unauthorized physical access	Big data analytics can offer the operators with the services like preventing unauthorized physical access in their network or subscribers own network. Big data analytics will let the operators know the usual activities and can generate alarm whenever some unusual happens on the network because of unauthorized access.	Location data, traffic data, network behavior data etc.	Capability of real time data analysis, capability of predictive analysis
	real time visibility on secured places	Operators can provide services like real time visibility in secured places like bank and shops with their network support and predict any kind of mismanagement though proper data analysis	Location data, historical data, live stream data	Capability of real time data analysis, capability of predictive analysis

	Device security monitoring	Use of a smartphone or another device, such as a laptop or tablet, as an unauthorized wireless hotspot to connect multiple users. This type of activity generates large volumes of data transmission and results in lost revenue for the provider. The system can identify the fraudulent or unintended usage that is violating a subscriber's wireless service agreement and the company's customer service team can then contact the subscriber to ask them to cease this type of activity or upgrade their contract	Location data, historical data, SDRs, device data, subscribers usage data	capability of big data storage, capability of predictive analysis, capability of data aggregation, capability of real time data analysis
Improving Customer care services	Reduce customer care calls	Big data analytics can help operators to recognize the service problem even for specific user concern, and solve it before the subscriber call the customer care for it. Or if they can predict any service failure beforehand they can let the subscribers know about it, it also will reduce the number of customer care calls	Subscriber detail record, Location data from the handset data, HLR, VLR data, operational data	Capability of real time data analysis to react immediately, capability of data aggregation, capability of predictive analysis to predict service failure
	Provide network information to the customer	With proper data analysis and real time dash boarding operators can offer the subscriber to follow their service or network or coverage in real time, this will let the subscriber to use data service or voice service in a proper area and also it will let the subscribers to compare the network coverage with other competitor operators	Subscriber location data, Network data from probes and other network equipment, coverage data	capability of data aggregation to aggregate the different types of data, capability of real time dash boarding to presenting the network information to the subscribers
	Improve intelligence of the customer care agents through real time visibility in customer experience	Big data analytics can improve and help to improve the customer care agents intelligence which will certainly earn the customer satisfaction through customer care calls	Network configuration data, SDR, location data	Capability of storing big data, capability of accessing databases with short latency, capability of real time data aggregation to help the customer care agents to learn the customer problem very quickly and provide solution in very less time
	Solve problem during the call	Most of the Operators are looking for this use case		Big data storage, real time data analysis, predictive analysis

	Automatically authorizing a customer care representative	Automatically authorizing a call center representative, who is speaking to a customer known to be having problems with their service, to present the customer with an offer that compensates them for their trouble and helps retain them as a customer. Examples of an offer could include a month of free service, a free device upgrade six months ahead of schedule and so on.	SDRs, customer location data, customer care representatives expertise data	Capability of big data storage, capability of real time decision making, capability of data aggregation
Transportation	Provide the traffic info to the customers	operators can provide services like providing the traffic info to the subscriber according to their location and requirements	Location data from the handset panels, traffic data from the governmental traffic info	Capability of data aggregation, capability of data segmentation, capability of predictive analysis, capability of real time analysis and real time reporting
	Real time traffic prediction and mobility pattern estimation	Operators can provide services like real time traffic prediction and mobility pattern estimation, subscribers can use services like these for journey planning	Location data from the handset data, traffic data from the governmental traffic info	Big data storage, predictive analysis, real time analysis
	Vehicle tracking for security purpose	For security purpose and criminal capturing, services like vehicle tracking can be offered by the operators. By proper data analysis operators can also find out that what kind of vehicle the subscriber is using	Location data from GPS, data from real time camera from the roads and highways, data from handset panels	capability of aggregating different types of data, real time data analysis, predictive analysis, capability of learning models
	Suggesting nearest gas stations	According to the location of the user and gas status of the car, operators can give some gas station suggestions. This will definitely impress the subscriber and reduce churn	Maps data from databases, location data from the handset data, gas status data from the sensors	capability of data aggregation, capability of real time data analysis
	Transportation recommendation and Route Mapping	In corporation with transport agencies or the government, mobile operators can provide transportation recommendation to the users. Operators can provides applications which can help the subscribers to map their route for journey	Location data from panels or GPS, traffic information from governmental server, maps from stored data base	Capability of storing big data, capability of real time data analysis, capability of predictive analysis to find out the easiest route, capability of data aggregation, capability of data real time reporting
	Managing vehicle fleets	Operators can provide services like managing vehicle fleets to the subscribers	Handover data, location data	Capability of real time data analysis, capability of data aggregation
	Tracking of driven kilometers of vehicles for taxation purpose	There is some kind of planning to change current car taxes based on driven kilometers, which will then be based on GPS tracking of cars	GPS data, location data from the handsets, taxation information data	Capability of Real time data collection, correlation analysis, statistical analysis etc.

Public Sector	Asset allocation	Military sectors can use mobile operators help to meet this use case, and operators can earn some revenue from the government	Location data, external data	Big data storage, predictive analysis
	Power grid information	Incorporation with government, operators can store and give power grid information to the subscribers	Location data, power grid info data from the government	Capability of data aggregation to aggregate the power info data and location of subscriber data, real time data analysis to provide real time info, predictive analysis to predict any power failure etc.
	Response to calamities	Weather fore casting and response to the calamities, alerting people is going to be one of the most important use cases of big data	Location data, weather forecasting data, historical data	Capability of real time data collection and analysis, capability of predictive analysis, capability of real time reporting
Business	Work as data center	This will bring the operators some revenues. As they are already handling lots of data and have the system for it, they can easily work as data centers for other small companies		Capability of Big data storage and analysis
	Deliver remote private cloud services	Operators can offer companies like heath care and financial services service like remote private cloud services where operators will host the server and the data delivery of the cloud service will be through their radio network.	Network data, traffic data, external data	Capability of Big data storage and analysis
	offer products In cooperation with retailers	By analyzing subscriber usage data operators can learn about the interest of the subscriber and in cooperation with retailers operators can provide advertising or offer their product to them which suits that subscribers interest	Subscriber detail records (SDR) data, users usage data from handset and HTTPs, sentimental data from the social networking sites	Capability of sentiment analysis to learn user sentiment about the retailers product and subscriber interest, capability of predictive analysis, capability of real time data analysis and reporting
	Reduce churn	Big data analysis will let the operators generate several new services according to the customers interest, will let the operators earn customer satisfaction through network and service improvement according to customer needs and these will reduce churn. Though proper data analysis operators can learn or predict which customers are about to churn and offer them new or better services by analysing their usage data, which will hello reducing churn	CDRs, SDRs, mobile positioning data form the user equipment, users usage data from handset panels, user behavior data from HTTPs, user sentiment data from social networking sites	Capability of big data storage, capability of data aggregation, capability of analyze unstructured data, retrospective analysis, capability of predictive analysis

	Increase ARPU	This is the ultimate business target for the operators and operators can get new ideas on increasing the ARPU by proper data analysis.	Sales data, user usage data, customer data, total revenue, number of users	Data aggregation, predictive analysis
	Increase sales	Big data analysis can help the operators to increase their product and service sales by providing special services through intelligent marketing	Sales data, billing data, market data	Big data storage, data aggregation, predictive analysis
	Reduce CAPEX and OPEX	Proper big data analysis can help the operators to reduce their total CAPEX and OPEX.	Sales data, operation data, product data, service data, other business data	Big data storage, data aggregation, predictive analysis
	Set the business target	Big data analysis can help the operators to set their business targets through market data analysis, business data analysis	Business data like sales, revenues, number of customers according to the location areas, historical data of previous years business, market research data like what the competitors are working on etc.	Capability of storing big data, capability of clustering, capability of pattern recognition, capability of predictive analysis
	Business target meet	Big data analysis can also help the operators to meet their business target. Proper data analysis can let the operators know if business is going down and what might be the reason and the solution for it.	Business data like sales, revenues, number of customers according to the location areas, historical data of previous years business, billing data, market research data like what the competitors are working on etc.	Capability of storing big data, capability of clustering, capability of pattern recognition, capability of predictive analysis
	Choose pre-loaded OTT services	Operators can do market research and subscriber data analysis to find out what are the most popular OTTs and can select those as pre-loaded service with their handsets	Market research data, OTT services data, subscribers sentiment data about those OTT services	capability of descriptive analysis, capability of predictive analysis, capability of finding correlation, capability of sentiment analysis
	Identifying next rising OTT services and offering those	operators can use their subscriber usage data, sentiment data to find out which might be the next rising OTT service and they can offer those to their subscriber	Market research data, OTT services data, subscribers sentiment data about OTT services	capability of descriptive analysis, capability of predictive analysis, capability of finding correlation, capability of sentiment analysis
	Selecting next best service, device, application	Using the stored data sets and analyzing those operators can select or predict their new best services, devices for network and application for the subscriber. This will raise the business	Market data, sales data	Descriptive analysis, predictive analysis
	Sell the data	Future use case, operators can sell their stored data to government or rapport agency or to some others. There are some challenges to meet this use case like privacy.		Big data storage

	Device recommendation	In corporation with device or hand set providers mobile operators can recommend the devices to the users and earn revenue from those companies	Device data, customer sentiment data about devices from the social networking sites, historical data about device sales	capability of predictive analysis, capability of insight generation, capability of sentiment analysis
	Generate own Applications	Operators can generate their own application analyzing subscriber interest and use those to attract consumers	subscriber sentiment data from social networks, subscriber usage data from handset and HTTPs	capability of predictive analysis, capability of sentiment analysis
	Improve decision making	Big data analysis can help the operators to improve their decision making skills by proper predictive analysis of proper data	According to use case	Big data storage, predictive analysis
	Customer relationship management	Big data analysis can let operators earn very good understanding on customer behavior, interest, sentiment, usage etc. Which will help the operators to maintain a very good customer relationship management	Subscriber usage data and behavior data from the handset panels and web browsing data, customer sentiment data from historical data from customer care calls, social networking sites, forums, blogs etc.	Capability of diverse data collection, capability of sentiment analysis, capability of predictive analysis, capability of data clustering
	Recognize and capitalize on next big streaming services	Like other services operators can predict next big streaming services which might bring them good revenue, and proper big data analysis can help operators to learn it.	Sales data, market data, subscribers usage data	Capability if storing big data, capability of data aggregation, capability of correlation, capability of predictive analysis
	generate new sources of revenue	Big data analysis can and is already providing the telecom industry several new ideas and opportunity to find out new sources of revenue	Business data like sales, revenues, service outputs, billing data, number of customers according to the location areas, historical data of previous years business, market research data like what the competitors are working on etc.	Capability of storing big data, capability of clustering, capability of pattern recognition, capability of predictive analysis
	Capture, retain and grow customers	Through big data analysis operators can reduce churn, capture new customers providing exciting services, offering new services which customers are interested in	Subscriber usage data and behavior data from the handset and HTTPs, subscriber sentiment data from the social networking sites, market data	Capability of diverse data collection, capability of sentiment analysis, capability of regression analysis, capability of predictive analysis, capability of data clustering
	Play a vital role building smarter cities	May be this is a marketing talk, but if technologies can build the so called smart city, mobile operators will play the most vital role with their stored big data and analysis	All kind of data operators handle	Capability of storing big data, capability of clustering, capability of pattern recognition, capability of predictive analysis, capability of descriptive analysis
	Enabling new business models	Big data analysis generates insight of data and operators may come up with new business models	Sales data, product data, other business data etc.	capability of big data storage, capability of insight generation from data analysis through predictive analysis, descriptive analysis etc.

	Predict up-selling potential	Through big data analysis operators can predict their upcoming selling potential and can optimize it accordingly	Sales data, product data, other business data etc.	Big data storage, predictive analysis
	create profitable new partnerships	Big data analysis can let the operators make some profitable partnerships with other retail or manufacture companies.	Device data, customer sentiment data from social media, customer behavior data	
	Use data as a service	Operators can offer companies and public sector bodies analytical insights based on real-time, location-based data service	mobile positioning data from handsets or GPS, customer sentiment data from social media etc.	Capability of big data storage, capability of insight generation, predicative analysis
Governmental	Security	Mobile operators can help government to maintain the security by providing different kind of supports	SDRs, CDRs, location data from HLR, VLR, GPS, sentiment analysis from the social networking sites, mobile terminal data,	capability of storing big data, capability of real time analysis, capability of predictive analysis, capability of clustering the data
	Criminal tracking	Mobile operators can help the government with their network data and subscriber data to find the criminals	SDRs, CDRs, location data from HLR, VLR, GPS, sentiment analysis from the social networking sites	capability of storing big data, capability of real time analysis, capability of predictive analysis, capability of clustering the data
	Crime tracking	Mobile operators can help the government with their network data and subscriber data to find the crimes	SDRs, CDRs, location data from HLR, VLR, GPS, sentiment analysis from the social networking sites	capability of storing big data, capability of real time analysis, capability of predictive analysis, capability of clustering the data
	improve first responder and emergency services	Governmental emergency services like hospital and fire brigade can improve their services by operators help. Through proper data analysis operators can provide services which can give these emergency service an predictive idea	Weather data, emergency services historical data, location data	capability of big data storage, capability of predictive analysis, capability of real time data analysis and real time reporting
	Public sentiment analysis	In corporation with mobile operators, government can learn the public sentiment about certain things and take decision accordingly	Subscriber usage data from handset panels, subscriber behavior and sentiment data from social networking sites	Capability of big data storage, capability of sentiment analysis, capability of text processing
	Transportation	Government can take help form the operators to improve the transportation service for public satisfaction	Location data, traffic info from government, mapping data etc.	Capability of accessing data bases in real time, real time data analysis, real time reporting capability, capability of data aggregation and predictive analysis
	healthcare	Operators can provide some healthcare services to improve		Big data storage, Descriptive analysis, Predictive analysis, real time analysis
	Online tax services	Operators can use big data analysis to provide services like online tax payment, reminding the subscriber about tax, providing tax calculation and earn revenue from the government	Tax information data from the government, SDRs, location data	Capability of big data storage, capability of descriptive analysis, capability of basic mathematical calculation

	Online permitting	Operators can provide services like online permitting, licensing from the government	Information data from the government, SDRs	Capability of big data storage, capability of in memory analysis
	Online form and applications	Operators can host all the governmental forms and application		Capability of big data storage, capability of real time reporting, capability of quick query from data bases
	Municipal courts and legal info	Operators can host governmental courts and legal info for the subscribers to access		Capability of big data storage, capability of real time reporting, capability of quick query from data bases
	Education	With the big data storage and analysis, operators can provide educational services for specific regions where much needed	Educational data, location data	Capability of storing big data, capability of real time reporting
Healthcare	Remote healthcare monitoring	Operators can provide services like remote healthcare monitoring where patients upload vital data and professionals access it and advice the patients accordingly	Health care data, patients data, location data,	Capability of big data storage, capability of data aggregation
	Connected hospitals	The operators can offer the hospitals to be connected to each other through their network and operators can provide their data center and data analysis procedures to them as rent.	External data	Capability of big data storage, data aggregation, correlation analysis
	Case conferencing	operators can store and share the case studies from different hospitals online to be accessed by others, or offer services like case conferencing	External data	Capability of real time data analysis, capability of storing big data
	Cloud computing for health	Operators can provide services like cloud computing and earn revenues from the healthcare	External data	Capability of Data aggregation, real time data analysis
	Certified Health data hosting and transferring	In health ecosystem operators can host and transfer health data	Certified health data from the health cares, location data, traffic data	Capability of big data storage, capability of clustering, capability of data transformation
	Drug authentication	Text messaging solution to ensure that the medicine is not counterfeit	Drug authentication data	Capability of big data storage, capability of clustering, capability of correlating, capability of real time data analysis, capability of text processing
	Emergency call center service improvement	Health care emergency call center services can be improved by the help of operators. Operators can offer special offers for the emergency call centers, or from the location data of the subscriber operators can suggest the nearest emergency call center number	Subscriber details record, location data, emergency call centers information data	Capability of big data storage, capability of data aggregation, capability of real time decision making

	Chronic disease management	Operators can provide services like chronic disease management where patients collects vital data and uploads it, then data processing like monitoring aggregation and orchestration can be done in the operators server. Then the service suggest an adviser or provided coach can review the data and give feedback	patients data like name, age, insulin level, blood info etc., historical data, adviser data etc.	Capability of store big data, capability of real time data analysis, capability of predictive analysis
	Emergency alerting and monitoring	Operators can provide services like emergency alerting and monitoring where in case of emergency, monitoring service notifies the family or friends and first responder gets the exact geo location, vital data is automatically sent to the primary care team and they collect the patient	Subscriber details record, contacts list, location data	Capability of big data storage, capability of data aggregation, capability of real time decision making
	Online payment service for healthcare	operators can provide online computerized payment services which can have numbers of healthcare professionals	patients data like name, age, address, bank info, health care center data like name, address, doctors name, bank info etc.	Consumer location data, payment data, professionals data
	Online medical library	Operators can host an online medical library, where they can provide a space for medical providers to store educational materials, instructional videos, and handouts for patients, other specialists, and family members.	Case studies, educational materials data, videos, patients´ historical data	capability of real time data analysis, capability of storing big data, capability of secure the data
Media and Entertainment	Sports advertisement	Through customer sentiment analysis operators can give customers favorite team updates. Operators can advertise on behalf of the sports teams to proper customers	SDRs, subscriber usage data from HTTPs, sentiment data from the social networking sites	Capability of sentiment analysis, capability of predictive analysis, capability of data segmentation
	pop up news service	Operators can know what kind of news a customer interested in and provide popup service to those news categories and also can earn some revenue.	SDRs, subscriber usage data from HTTPs, subscriber behavior data from the handset panels and social networking	Capability of big data storage, capability of subscriber behavior analysis, capability of diverse data collection, capability of real time data analysis
	TV programs advertisement	In association with specific TV channel, with the big data analysis operators can recommend its user the TV programs they might like from that specific TV channel. This will bring some revenue from the TV channel.	SDRs, location data, subscriber usage and sentiment data from social networking sites, TV program database	Capability of sentiment analysis, capability of pattern recognition, capability of clustering the data sets, capability of real time data analysis
	Tracking of how consumers read/consume different kind of news	Big data analytics will allow the operators to know how their subscribers read/	User usage data , location data from handsets	Capability of real time data collection and analysis

Product and services	Real time performance analysis	Big data analysis can let the operators analyze the performance of their services in real time, which will help the operators to work on it if required instantly	Service data from network elements, probes, CDRs, Handover data	Capability of diverse data collection, capability of data aggregation, real time analysis
	Pricing impact and stimulation	Through proper data analysis operators can be able to predict or learn new pricing impact in business and on customer sentiment as well and stimulate accordingly	Pricing data, billing data, market data, customer sentiment data	Capability of diverse data collection, capability of data aggregation, real time analysis, capability of sentiment analysis
	Cannibalization impact of new services	Impact of new services from existing services or upgrade of existing services can be predicted through proper data analysis and proper steps can be taken accordingly	Service level data , market research data, subscribers requirement or interest data from surveys	Capability of retrospective analysis, predictive analysis, big data storage, clustering
	What-if analysis for new product launch	What-if analysis for new product launch can help the operators to learn about how the new products will have impact and market or how it can be optimized	Product data, market research data, previous product data	Capability of big data storage, predictive analysis, retrospective analysis, clustering
	Service assurance	Service level agreement (SLA) management, Service problem management, Service quality management	Service level data from CDRs, probes, network element interfaces, agreement data	capability of storing big data, capability of retrospective analysis, capability of data aggregation
	Product and service development	Proper data analysis can provide the operators or vendors new idea on product and service development. By analyzing pre-sale product and service buzz, post-sale satisfaction can help operators for product or service development	Product data, market data, service data	Capability of data aggregation, capability of predictive analysis
	Improving supply chain management	Big data analytics allows the operators to improve their supply chain management	Supply chain data, product data, sales data, subscriber location data etc.	Capability of storing big data, capability of retrospective analysis, capability of data aggregation
Billing	Bill shock prevention	Operators can use their big data and big data analytics capability to warn the subscribers' about their bill time to time	Billing data, subscribers' usage data, CDRs	Capability of real-time data aggregation, capability of Big data storage
	Accurate billing	Operators can ensure accurate billing and also identify bill frauds by big data analytics	Billing data, CDRs, XDRs, Subscribers' usage data	Capability of big data storage, capability of real-time data analysis
Others	Banking, Insurance, financial services, partner analysis, cost and contribution analysis etc.	Operators can utilize their big data analytics capability to generate other use cases.	According to use case	According to use case

Table 22: Potential big data use cases in telecom Industry

Appendix 4A

Preprocessing sub tasks and techniques

Binning methods smooth a sorted data value by consulting values around it and the sorted values are distributed in number of bins. Binning methods can also be called as local smoothing as it consults with the values around it or "neighborhood" (Han, et al., 2011). Smoothing by binning can be done in three different ways, they are: Smoothing by bin means, where each value in bin is replaced by mean value of the bin; Smoothing by bin medians, where each value in bin is replaced by median value of the bin; and Smoothing by bin boundaries, where the minimum and the maximum value in given bin are addressed as the bin boundaries.

Regression is a process of smoothing data by fitting the data to a function. Linear regression, multiple linear regression are examples of two different type regression. Linear regression is about finding the best line to fit two attributes so that one attribute can be used to predict the other one, and multiple linear regression is an extension of its which involves more than two attributes.

Clustering is one important process of outlier detection. To handle noisy data or find out inaccurate value clustering process is one best choice. In clustering method the similar or closer values are organized in groups and values that fall outside of the set of clusters may be considered as outliers. Clustering is very useful method for redundancy elimination, outlier detection.

Smoothing the data is performed to remove noise from the data by binning, regression, and clustering. Smoothing is also a form of data cleaning as described in Data Cleaning section. Smoothing the data, which is quite sensitive to data reliability, may allow the reduction of measuring time in experiments such as diffraction (Han, et al., 2011).

Data Aggregation is applying summary or aggregation operations to the data. Aggregation is typically used for constructing data cubes for analysis if the data at multiple granularities. Data aggregation aggregates data from different sources or data sets from the same source for proper data analysis and insight generation.

Data Generalization is the process of replacing low-level or raw data by high-level concepts using concept hierarchies. For example, attributes like street can be replaced by city if required.

Data Normalization is done to scale data attributes within a small specified range. Sometimes data normalization is very important and sometimes it is not performed at all, it totally depends on the requirements.

Attribute construction is a very important part of data transformation and technique of data preprocessing. In attribute construction new attributes are constructed from the available attributes to help the mining process.

Data cube aggregation generates data cubes out of the data set when applied, reduces the data and analysis becomes easier and less time consuming.

Attribute subset selection is another process of data reduction which identifies the irrelevant, weakly relevant or redundant attributes and removes them.

Dimensionality reduction is the process where encoding of data is done to reduce the data and then the encoded data is used for data presentation. The encoded data can be decoded any time if required.

In **Numerosity reduction**, data are replaced by alternative, smaller data representations as parametric models.

Another data reduction procedure is **Discretization and concept hierarchy generation**. In this process raw data values for attributes are replaced by ranges or higher conceptual levels. Discretization is a process of generating concept hierarchy. This whole process can be divided into three parts naming unsupervised discretization, supervised discretization and generating the concept hierarchy. Any one of discretization process is applied from supervised and unsupervised to generate the concept hierarchy.

Data Filtering:

Data filtering is one very important data transformation technique. Data filtering actually covers a broad area, at one end of spectrum it deals with simple problems like corrupt data and at another end it deals with noisy data. The ideal filtering technique removes the irrelevant features with minimal distortion of the relevant signal features in time domain, frequency domain or time-frequency domain. In time domain filtering the mean or median of the measured data in a window of predetermined size is taken, in frequency domain filtering data is transformed via Fourier analysis and high frequency contributors are eliminated, and in time-frequency domain filtering the measured data is transformed in the time and frequency domain simultaneously.

Data Ordering:

The main objective of data ordering is to organize data in proper location for further retrieval and analysis; it is most applicable when they are stored in relational or network database management systems. Data ordering involves conceptual model preparation, entity identification, labeling the attributes within the entities etc.

Data Editing:

When there are unstructured data then data editing requires the most. Data editing helps converting the unstructured data to structured data or semi-structured data at some extent. When data consists of text or symbols or strings of characters which bear unique information, data editing is converts those symbols into information. Data editing is a very intense work as incorrect editing may bring out wrong information.

Noise Modeling:

Another important division of data transformation technique is noise modeling. As mentioned earlier that there are several reasons why data gets noisy and noise modeling helps removing the noise or sometimes replacing with proper information.

Fourier transform is one most common noise modeling for data preprocessing (Famili, et al., 1997). There are several more adaptive schemes for noise estimation. Bayesian, Maximum likelihood,

Correlation and Covariance matching are few common noise estimation procedures. Another form of noise modeling and smoothing is compression. Data compression can enhance and improve interpolation which results in better classification on the testing data sets (Famili, et al., 1997).

Data Visualization:

Data visualization is the process of visual representation of data, in other words, data visualization is visual representation of information that has been abstracted in some schematic form, including attributes or variables for the units of information.

According to Friedman (2008) "the main goal of data visualization is to communicate information clearly and effectively through graphical means. It doesn't mean that data visualization needs to look boring to be functional or extremely sophisticated to look beautiful. To convey ideas effectively, both aesthetic from and functionality need to go hand in hand, providing insights into a rather sparse and complex data set by communicating its key-aspects in a more intuitive way. Yet designers often fail to achieve a balance between form and function, creating gorgeous data visualization which fail to serve their main purpose which is to communicate information" (Friedman, 2008).

Data Elimination:

Data elimination as preprocessing technique is performed to achieve typically two objectives (Famili, et al., 1997):

- The volume of the data is reduced substantially
- The data is partially classified

Data Selection:

Data selection is one of the effective ways of solving 'large amounts of data' problem. Data selection has numbers of advantages, and several researchers have deployed methods for analyzing and categorizing data in much smaller data sets. In (Kelly & White, 1993) the authors proposed a clustering technique to analyze large amounts of image data, where the original technique is to represent the image using numbers of small pixel value. Data selection makes the data analysis process easier and makes preprocessing technique smarter.

Principal Component Analysis:

Lots of researcher and industry experts consider Principal Component Analysis (PCA) as most important technique of data preprocessing. The main foal of performing PCA is to selecting the proper attributes for data analysis.

PCA uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variable called principal components (PC). PCA is performed in such a way that the first PC has the largest possible variance and each succeeding components in turn has the highest variance possible under the constraint that it be orthogonal to the preceding components.

Data Sampling:

Sometimes for data analysis the data set needs to be split into parts. Training, testing, evaluation the performance of the data analysis requires this kind of actions and data sampling is the best way to do it. In data sampling the most important issue is to make sure that the correct choices of samples are made when sampling the data otherwise the evaluation result will be unsatisfactory.

Adding new Features:

Constructive induction manually or automatically derives new features in term of the existing ones. Data-Driven, Knowledge-Driven and Hypothesis-Driven are the three different procedures deriving new features following constructive induction procedure (Famili, et al., 1997).

Rule based fuzzy systems are another procedure of adding new features which is typically formulated as If-Then statement where "If" part is called the premise and "Then" part builds the conclusion (Famili, et al., 1997).

Time Series Analysis:

Time series analysis means transforming data into a static collection of features that represents a view of the operation at some time. Time series analysis is a reliable process when data exhibits too much variation or non-stationary behavior. Because of process equipment, raw material usage, human interventions etc. data got variations and time series analysis is the approach to deal with this problem.

Data Fusion:

Data fusion is the process of integration of multiple data and knowledge representing the same real world object into a consistent, accurate, and useful presentation. There are three different types of data fusion naming low, intermediate and high depending on the processing stage at which it takes place.

Data Simulation:

Data simulation is the technique of preprocessing which deals with the problem of unavailable or immeasurable parameters in large measurement spaces. It might be possible to simulate those unavailable parameters and apply them into the entire measurement space so that the effects of these parameters can be investigated if required. Data simulation overlaps with knowledge-Driven constructive Induction when the goal of induction is to induce decision trees.

Dimension Analysis:

Dimension analysis is the practice of checking relations among attributes by identifying their dimensions. Dimension analysis generates qualitative rather than qualitative relationships which makes it more reliable process. The goal of using dimension analysis is to transform the existing measurement space into a series of dimensionless terms that can be sued for data analysis (Famili, et al., 1997).

Appendix 5A

Function	Short description	Scoring (0-3)	Score requirements
Missing value analysis	Missing value analysis helps address several concerns caused by incomplete data. If cases with missing values are systematically different from cases without missing values, the results can be misleading. Also, missing data may reduce the precision of calculated statistics because there is less information than originally planned. Another concern is that the assumptions behind many statistical procedures are based on complete cases, and missing values can complicate the theory required.	3	capability to Impute, calculate or predict missing value
		2	Capability to fill up missing values by attribute mean, median etc.
		1	Capability to fill up missing values by some constants like zero or infinity
		0	No capability to handle missing value
Filtering	Filters works with records or rows and columns of data in the database. The conditions that are set are compared with one or more fields in the record. If the conditions are met, the record is displayed. If the conditions are not met, the record is filtered out so that it isn't displayed with the rest of the data records. Filtering does not permanently remove records it just temporary hides them from view.	3	Capability of filtering attributes, horizontal filtering, vertical filtering
		2	Capability of attribute filtering
		1	Capability of rows filtering
		0	No capability of filtering
Aggregation	The process of redefining data into a summarization based on some rules or criteria.	3	Data aggregation of different types from diverse sources
		2	Data aggregation of different types from single source
		1	Very simple data aggregation capability
		0	No capability of data aggregation
Validation	Data validation is the process of ensuring that a program operates on clean, correct and useful data. It uses routines, often called "validation rules" or "check routines", that check for correctness, meaningfulness, and security of data that are input to the system. The rules may be implemented through the automated facilities of a data dictionary, or by the inclusion of explicit application program validation logic	3	Capability of data validating with several validation rules like "cross-validation", "split-validation", "Bootstrapping validation"
		2	Capability of data validation for some specific application like XSD validation, credit card validation, data type validation etc.
		1	Capability of form level, field level, data saving, and range validation
		0	No capability of data validation
Correlation	Dependence refers to any statistical relationship between two random variables or two sets of data. Correlation refers to any of a broad class of statistical relationships involving dependence. Correlation is a statistical technique that can show whether and how strongly pairs of attributes are related.	3	Capability of calculating correlation matrix
		2	Capability of calculating correlational coefficients
		1	NA
		0	No capability of data correlation calculation
Enrichment	Data enrichment is a general term that refers to processes used to enhance, refine or otherwise improve raw data	3	Use of fuzzy logic to assist a search activity; accessing related data from other sources and bringing the data into a single virtual (for example, providing a link) or physical location; and, correcting misspellings
		2	Capability of enrichment where external data from multiple sources is added to the existing data set to enhance the quality and richness of the data
		1	Very basic data enrichment capability like reduction, addition of redundancy etc.
		0	No capability of data enrichment

Data profiling	Data profiling, also called data archeology, is the statistical analysis and assessment of the quality of data values within a data set for consistency, uniqueness and logic.	3	Capability to Utilize different kinds of descriptive statistics such as minimum, maximum, mean, mode, percentile, standard deviation, frequency, and variation as well as other aggregates such as count and sum. Additional metadata information obtained during data profiling could be data type, length, discrete values, and uniqueness, occurrence of null values, typical string patterns, and abstract type recognition. The metadata can then be used to discover problems such as illegal values, misspelling, missing values, varying value representation, and duplicates.
		2	Capability of single columns profiling individually to get an understanding of frequency distribution of different values, type, and use of each column. Embedded value dependencies can be exposed in cross-columns analysis
		1	Basic data profiling tasks like statistical calculation and validation
		0	No capability of doing data profiling
Outlier detection and analysis	Outlier detection is the search for data items in a dataset which do not conform to an expected pattern	3	Detect outlier based on distance, based on data density, based on local outlier factors, based on class outlier factors
		2	Z-score method, Box-plot method outlier analysis
		1	Simple supervised, unsupervised and semi-supervised outlier detection
		0	No capability of outlier detection
Meta data transformation	data transformation converts a set of data values from the data format of a source data system into the data format of a destination data system	3	Meta data aggregation, modification, sorting, conversion etc.
		2	Metadata transformation at a small extent
		1	Metadata information generation
		0	No metadata transofrmation capability
Sampling	The idea behind data sampling is commonplace in any statistical analysis: in order to get results faster, you analyze a sub-set of data to identify trends and extrapolate aggregate results based on the percentage of overall traffic represented in the sub-set.	3	Capability of doing satisfied sampling, bootstrapping sampling, model-based sampling, kennard-stone sampling, split data sampling
		2	Capability of doing one or two among mentioned above
		1	NA
		0	No capability of doing data sampling
Discretization	Discretization refers to the process of converting or partitioning continuous attributes, features or variables to discretized or nominal attributes/features/variables/intervals	3	Discretize by frequency, discretize by size, discretize by binning, discretize by user specification, discretize by entropy
		2	Discretization by few of the mentioned above
		1	NA
		0	No capabilities of data discretization
Clusterring	Clustering is the task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense or another) to each other than to those in other groups	3	Connectivity based clustering, centroid based clustering, distribution based clustering, and density based

	(clusters)		clustering. And capability of clustering performance measurement.
		2	Only clustering capability, no performance measurement capability.
		1	Only segmentation
		0	No capability of clustering
Transformation	Transformation of data from one node to another in a simplest and accurate way	3	Accurate data transformation capability
		2	NA
		1	NA
		0	No capability of data transformation
Reduction	Data reduction is the transformation of numerical or alphabetical digital information derived empirically or experimentally into a corrected, ordered, and simplified form. The basic concept is the reduction of multitudinous amounts of data down to the meaningful parts	3	Remove correlated attributes, remove useless attributes, remove attribute range, backward elimination
		2	Data reduction at some extent
		1	Very weak data reduction capability
		0	No capability of data reduction
Attribute name, value and role modification	Change the name and the role of the attributes	3	Name, role, value modification of the attribute
		2	Name and role or name and value modification
		1	Only name modification or renaming
		0	No capability of value, name or role modification of the attributes
Optimization	Data optimizations is most commonly known to be a non-specific technique used by several applications in fetching data from a data sources so that the data could use in data view tools and applications	3	Attribute selection, attribute generation, feature selection, feature generation, Optimize by generation, optimize selection, ID generation, weight generation etc.
		2	Attribute and feature selection
		1	NA
		0	No capability of optimization
Attribute generation	Constructs new user defined attributes using mathematical expressions.	3	Capability of generating attribute using basic, log and exponential, trigonometric, statistical, text, date, process, etc expressions
		2	Capability of generating attribute using some basic expressions
		1	NA
		0	No capability to generate new attribute
Sorting	Sorting is the process of arranging data into meaningful order so that you can analyze it more effectively	3	Sort, shuffle, sort by pareto rank etc.
		2	Only simple sorting and shuffle
		1	only sorting
		0	No capability of sorting
Rotation	Rotation of example set when required	3	Pivot, de-pivot, transpose
		2	Pivot, de-pivot
		1	Only transpose

		0	No capability of rotation
Set operations	Set operations refer to query operations that produce a result set that is based on the presence or absence of equivalent elements within the same or separate collections (or sets)	3	Join, append, set minus, intersect, union, superset, Cartesian product etc
		2	capability of doing Some of the mentioned above
		1	capability of doing very few of them
		0	No capability of set operations
Clasification and regression	Data classification is the categorization of data for its most effective and efficient use. Data can be classified according to any criteria, not only relative importance or frequency of use. For example, data can be broken down according to its topical content, file type, operating platform, average file size in megabytes or gigabytes, when it was created, when it was last accessed or modified, which person or department last accessed or modified it, and which personnel or departments use it the most.	3	Lazy modeling, Bayesian modeling, Tree induction, Rule induction, Neural Net training
		2	Capability of doing some of the above mentioned
		1	Capability of doing very few of the above mentioned
		0	No capability of data classification
Regression analysis	Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables	3	Function Fitting, Logistic regression, support vector modeling, discriminant analysis, meta modeling
		2	Capability of doing some of the above mentioned
		1	Capability of doing very few of the above mentioned
		0	No capability of regression
Data manipulation	Data manipulation is the process of taking data and manipulating it in a method to be easier read or organized	3	Data manipulation by column, row, matrix etc. With auto binner, numeric binner, CAIM binner etc.
		2	Data manipulation at some extent
		1	NA
		0	No capability of data manipulation
Principal component analysis	Principal component analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components	3	Principal component analysis, independent component analysis, Generalized Hebbian algorithm
		2	NA
		1	NA
		0	NO capability of principal component analysis
Type conversion	Type conversion, typecasting, and coercion are different ways of, implicitly or explicitly, changing an entity of one data type into another	3	Nominal to Numerical, Numerical to Binominal, Nominal to Binominal, Numerical to Polynomial, Numerical to Real, Real to Integer, Nominal to Text, Nominal to Date, Text to Nominal, Date to Numerical, Date to Nominal, Guess Types
		2	Capability of doing some of the conversions mentioned above
		1	Capability of doing very few of the conversions mentioned above
		0	No capability of data type conversion

Attribute weighting	Weighting several attributes according to correlation, rule, value etc.	3	Weighting by Correlation, Weighting by Information Gain, Weighting by Rule, Weighting by Deviation, Weighting by Chi Squared, Weighting by Gini Index, Weighting by Tree Importance, Weighting by Uncertainty, Weight by Relief, Weighting by SVM, Weighting by PCA, Weighting by Component, Weighting by User Specification
		2	Capability of doing some of the weighting mentioned above
		1	Capability of doing very few of the weighting mentioned above
		0	No capability of weighting

Table 23: Scoring conditions of the features

Appendix 5B

Tools	RapidMiner	Pentaho Data integrator	Data Preparator	Talend open studio for data integration	Clover ETL community edition	Orange	Excel	Matlab	TANAGRA	KNIME	SPSS Statistics	R	Weka
Features													
Missing value analysis	3	3	2	2	2	3	3	3	2	3	3	3	2
Filtering	3	2	3	3	2	3	2	2	2	3	2	3	3
Aggregation	3	3	2	3	2	3	3	2	1	3	3	2	2
Validation	2	2	1	3	2	0	2	2	2	3	3	3	2
Correlation analysis	3	2	0	3	2	2	2	2	3	3	2	3	3
Enrichment	3	2	2	3	2	2	2	2	1	3	3	3	2
Data synchonization	3	3	1	2	2	2	2	2	1	2	2	3	2
Profiling	3	2	2	2	2	3	2	2	1	3	3	3	3
Outlier detection and analysis	2	2	2	3	1	3	2	3	2	3	3	2	3
Meta data transformation	3	2	2	3	2	2	2	2	2	3	3	3	2
Sampling	3	2	2	2	2	3	3	3	2	3	3	2	3
Discretization	2	0	2	2	2	3	2	2	0	3	2	3	2
Clustering	3	2	2	2	1	2	2	3	3	3	2	2	3
Transformation	3	3	3	3	3	3	3	3	3	3	3	3	3
Reduction	2	2	2	2	2	1	3	3	0	3	2	2	2
Name, role and value modification	3	2	2	3	2	3	2	2	1	3	2	2	2
Type conversion	3	2	2	2	2	2	2	2	1	3	3	2	2
Optimization	3	2	2	2	2	3	3	2	2	3	2	3	2
Attribure generation	3	3	2	3	3	3	3	2	0	3	3	2	2
Sorting	2	3	2	2	2	2	2	3	2	2	3	3	2
Rotation	3	3	0	3	2	1	3	2	0	3	2	3	2
Set operations	3	3	3	2	3	3	3	3	3	2	3	3	2
Classification	3	1	0	2	0	3	3	2	1	3	2	3	3
Regression	3	0	0	2	0	3	2	2	3	3	1	3	3
Manipulation	3	2	2	2	2	2	2	3	2	2	2	3	2
PCA	3	0	0	0	0	3	3	3	0	3	3	3	3
total	73	53	43	61	47	63	63	62	40	74	65	70	62

Table 24: Scoring of tools according to preprocessing features capability

Appendix 5C

Tools		RapidMiner	Pentaho Data integrator	Data Preparator	Talend open studio for data integration	Clover ETL community edition	Orange	Excel	Matlab	TANAGRA	KNIME	SPSS Statistics	R	Weka
Feature domains	Features													
Ease of use	Self Organizing	x	x	x	x	x	x	x	x	x	x	x	x	x
	Flexibility	x	x	x	x	x	x	x	x	x	x	x		x
	numbers of options/features	x			x						x	x		x
	User friendly GUI	x	x	x	x	x	x			x	x			x
	Automation	x	x		x	x	x		x		x		x	x
	Easy to configure	x	x	x	x	x	x	x		x	x	x	x	x
	Easy upgrate	x	x		x	x	x			x	x		x	x
	Groovy Editor	x	x	x	x	x	x			x	x			x
	Query language	x	x		x	x	x		x		x	x	x	
	scripting language	x	x		x	x	x	x	x		x	x	x	
	process storage	x	x	x	x	x					x			x
	Process import/export	x	x		x	x	x		x	x	x	x	x	x
	Minimun Coding requirement	x	x	x	x	x		x		x	x	x		x
Performance	Low Latency	x	x	x		x	x	x	x		x	x	x	x
	Availability	x	x	x	x	x	x	x	x	x	x	x	x	x
	In memory preprocess	x	x	x	x	x	x	x	x	x	x	x	x	x
	Smart Exprot and Import of data	x	x		x	x			x		x	x	x	x
	real time Preprocess	x	x		x	x		x	x		x	x	x	x
	ELT support				x									
	Reliable	x	x	x	x	x	x	x	x	x	x	x	x	x
	Protect data loss	x	x		x	x	x	x	x	x	x	x	x	x
	Less memory consumption		x	x		x	x	x		x	x	x	x	x
	Accurate	x	x		x	x	x	x	x	x	x	x	x	x
Error Management	Error detection	x	x		x	x			x		x		x	x
	Auto fixing	x							x		x		x	
Price	Free	x	x	x	x	x	x			x	x		x	x
	Low price							x	x			x		

Table 25: Tools list with performance and usability features

Appendix 5D

Tools		RapidMiner	Pentaho Data integrator	Data Preparator	Talend open studio for data integration	Clover ETL community edition	Orange	Excel	Matlab	TANAGRA	KNIME	SPSS Statistics	R	Weka
Feature domains	Features													
Data connectivity	Database connectivity	x	x		x	x	x	x	x	x	x	x	x	x
	Access to all data sources	x	x		x	x	x	x	x		x	x	x	
	Access to all data types	x							x		x		x	
	Data extraction	x	x		x	x	x	x	x	x	x	x	x	x
	Multi-dimensional format data delevery	x	x		x	x			x		x	x	x	x
	NoSQL support	x	x		x	x	x				x		x	
	Hadoop extension	x	x		x	x					x		x	x
	High availability of Data	x	x	x	x	x		x	x	x	x	x	x	x
Advance	Parallel processing		x		x	x		x	x		x		x	
	Insight generation	x	x	x	x	x	x		x		x	x	x	x
	Modelling	x		x			x	x	x	x	x	x	x	x
	Series analysis	x					x	x			x	x	x	x
Text Analysis	Signs analysis	x									x		x	
	Exploration	x					x				x		x	x
	Migration	x					x				x		x	x
	Tokenize	x					x	x	x		x		x	x
	Rromve stop words	x					x		x		x		x	x
	Box of word creation	x					x	x	x		x		x	x
	Steming	x					x				x		x	x
Data Visualization	Multi-dimensional plotting	x	x		x	x	x	x	x		x	x	x	x
	Documentation	x			x	x	x		x	x	x	x		x
	Dashbording	x	x				x	x	x		x	x		x

Table 26: Tools list with analytics features

Appendix 5E

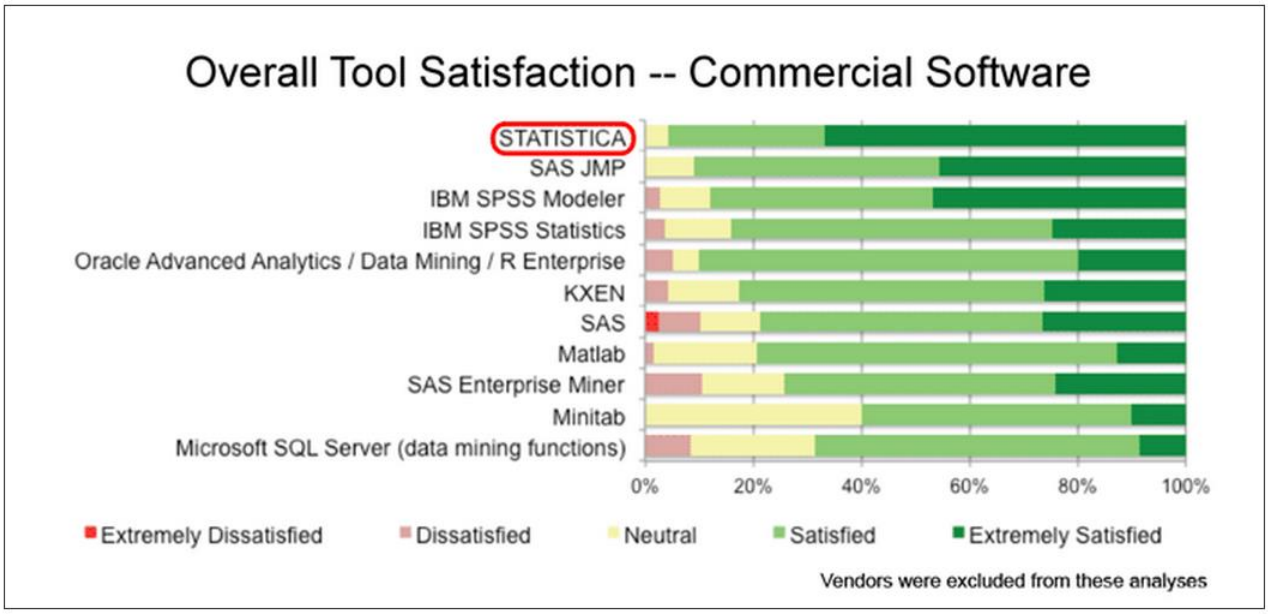

Figure 22: Commercial tools comparison (1) published in 'Data miner survey by Rexer Analytics' (Statsoft, 2013)

Appendix 5F

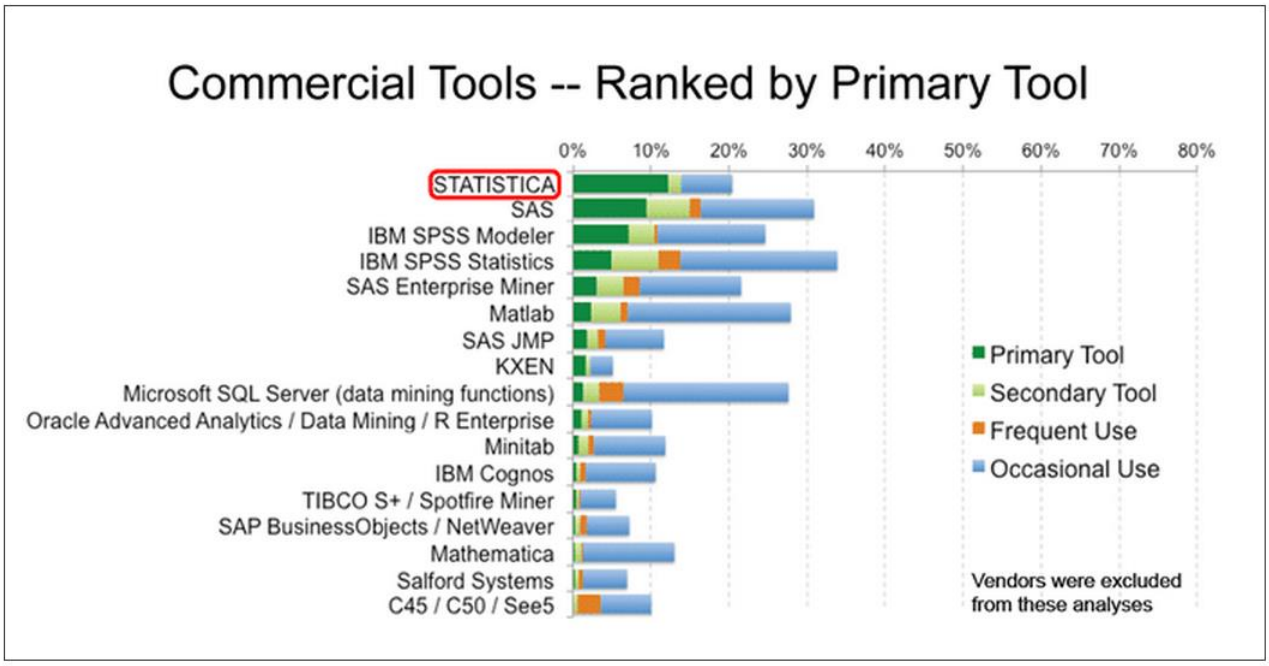

Figure 23: Commercial tools comparison (2) published in 'Data miner survey by Rexer Analytics' (Statsoft, 2013)

Appendix 6A: RapidMiner

Task 1

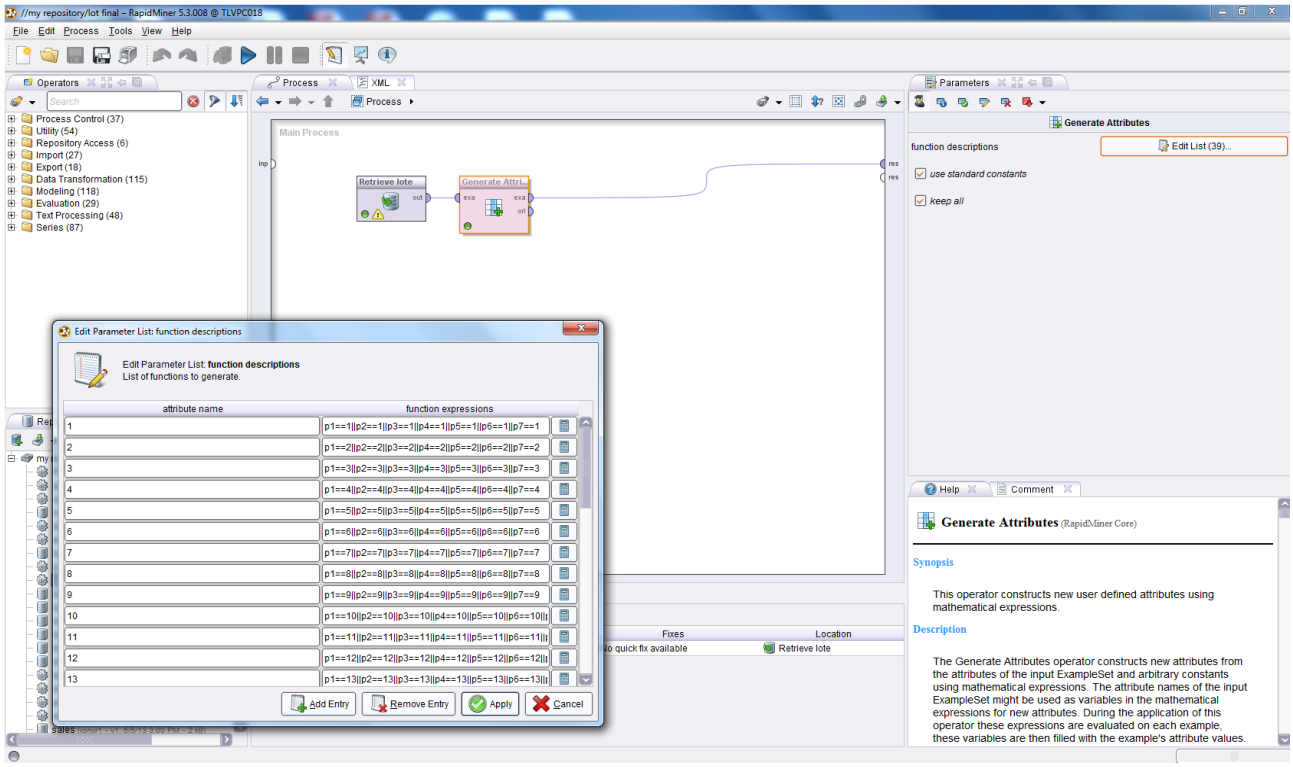

Figure 24: Preprocessing task 1 on RapidMiner

Task 2

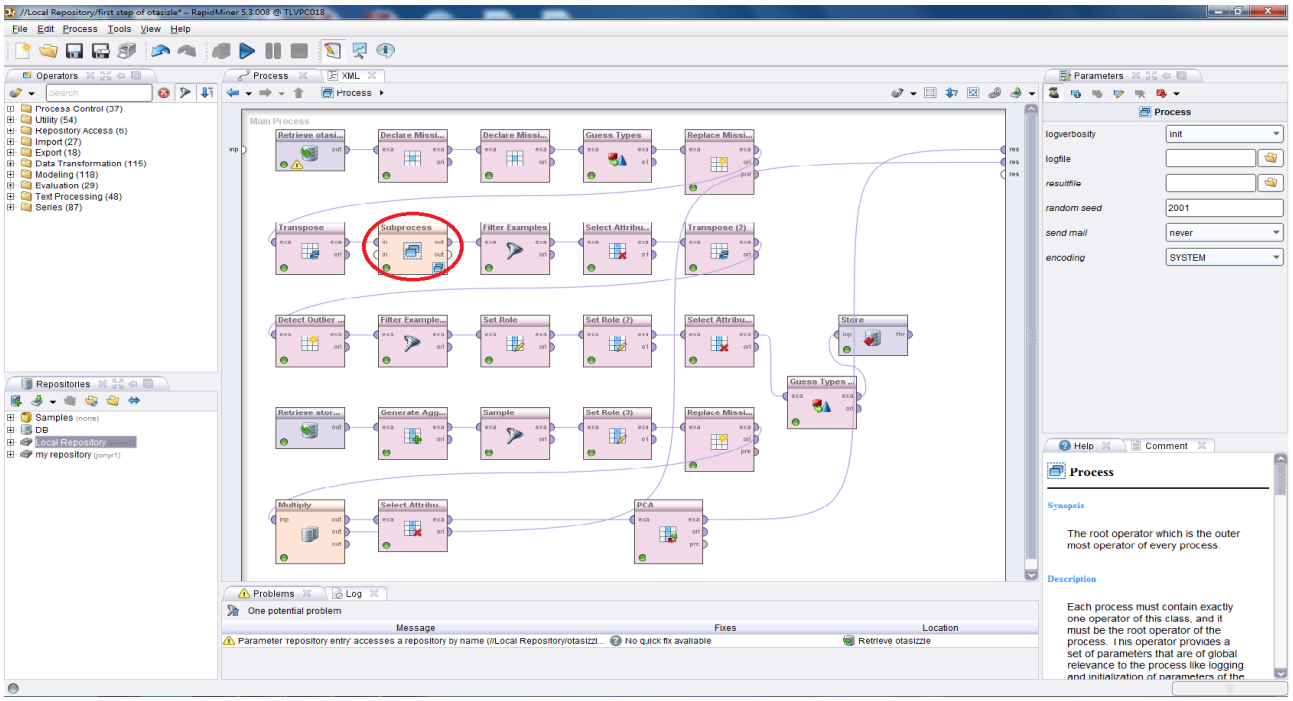

Figure 25: Preprocessing task 2 on RapidMiner

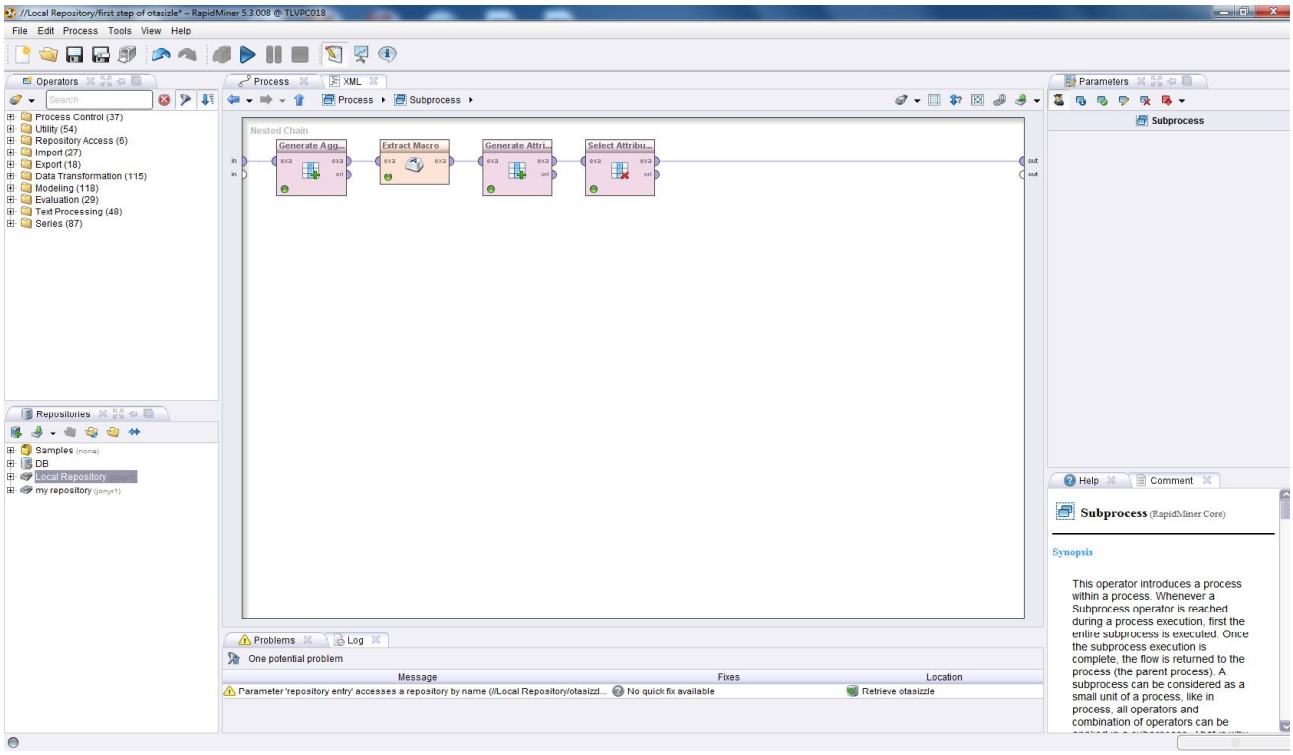

Figure 26: Sub-process of preprocessing task 2 on RapidMiner

Task 3

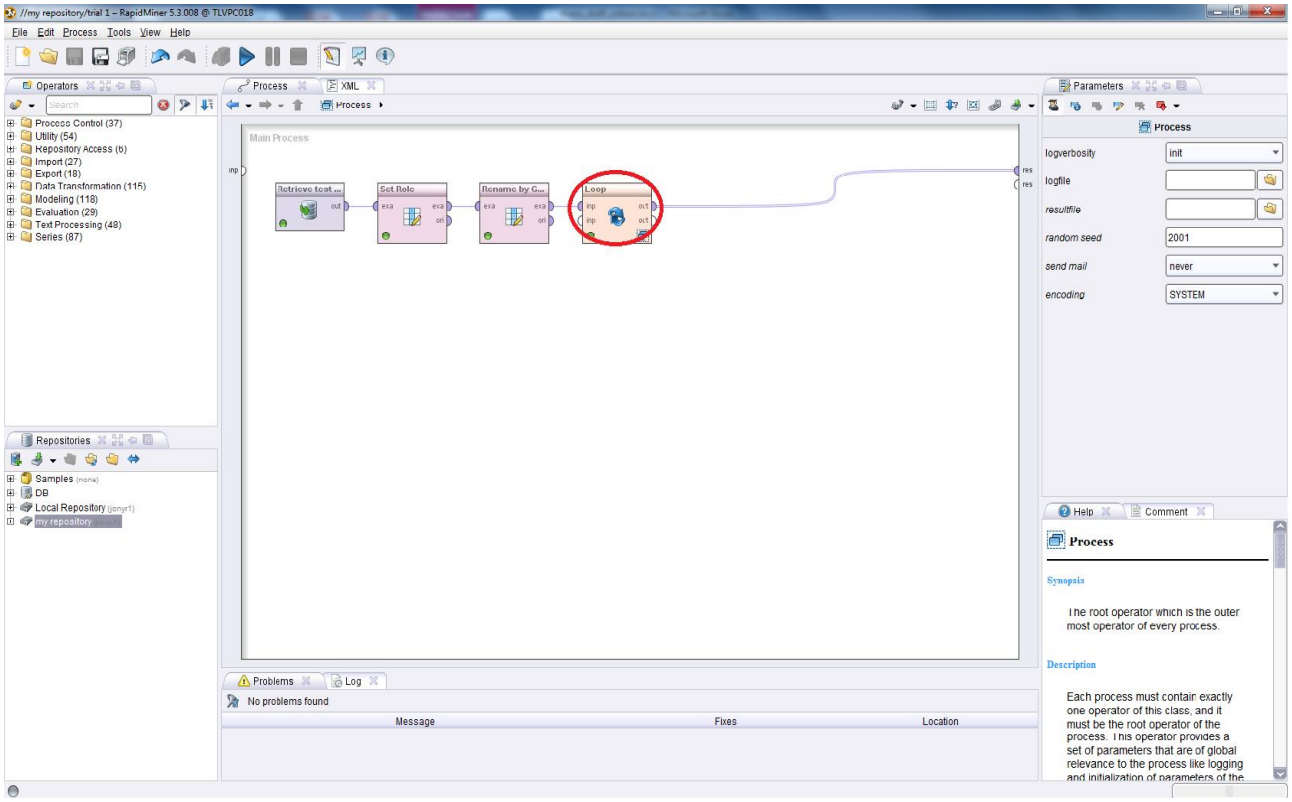

Figure 27: Preprocessing task 3 on RapidMiner

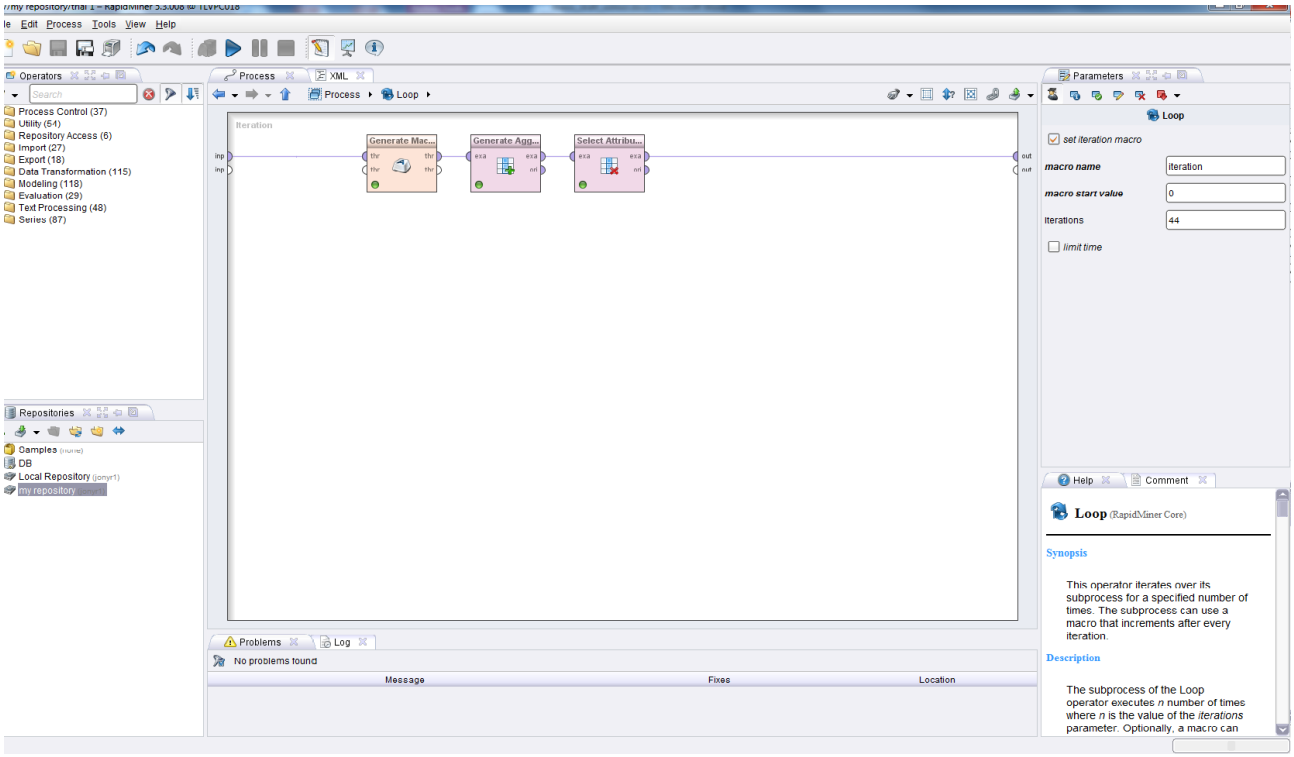

Figure 28: Sub-process of preprocessing task 3 on RapidMiner

Task 4

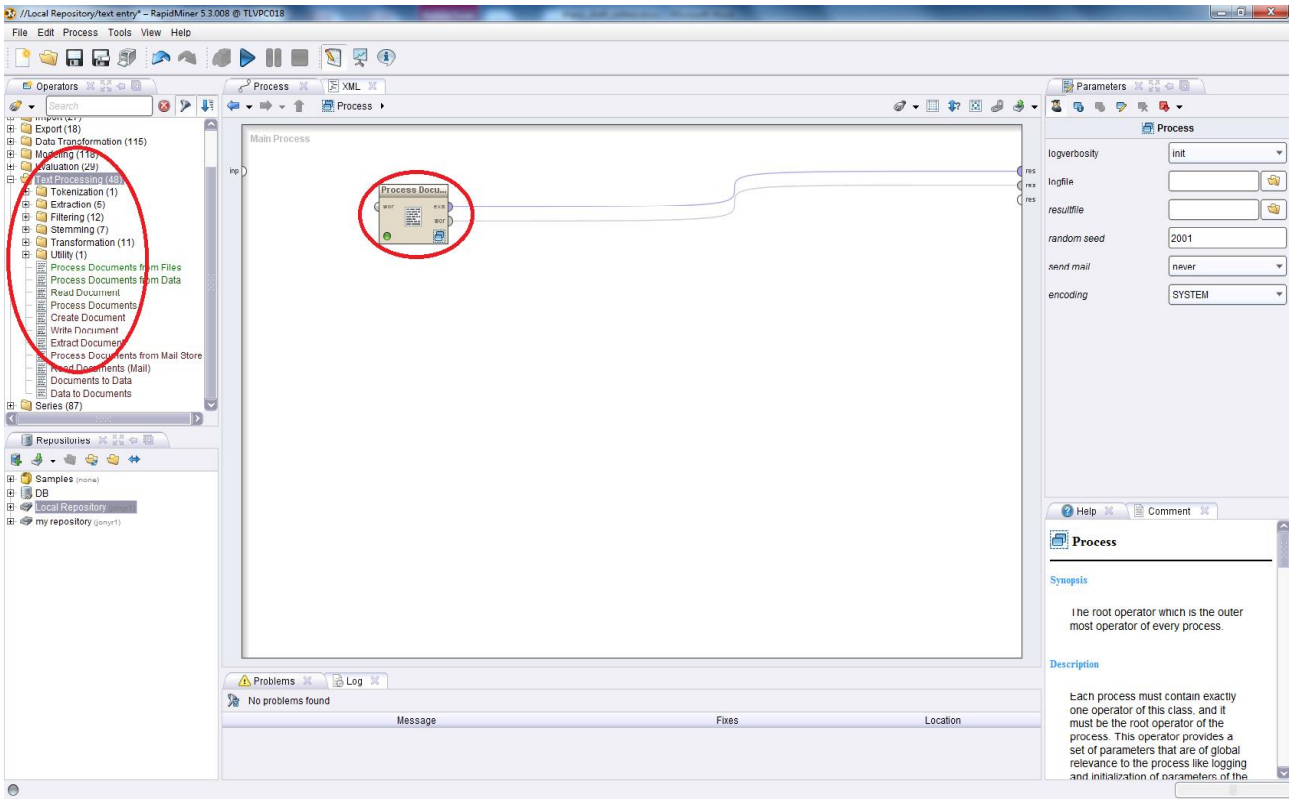

Figure 29: Preprocessing task 4 on RapidMiner

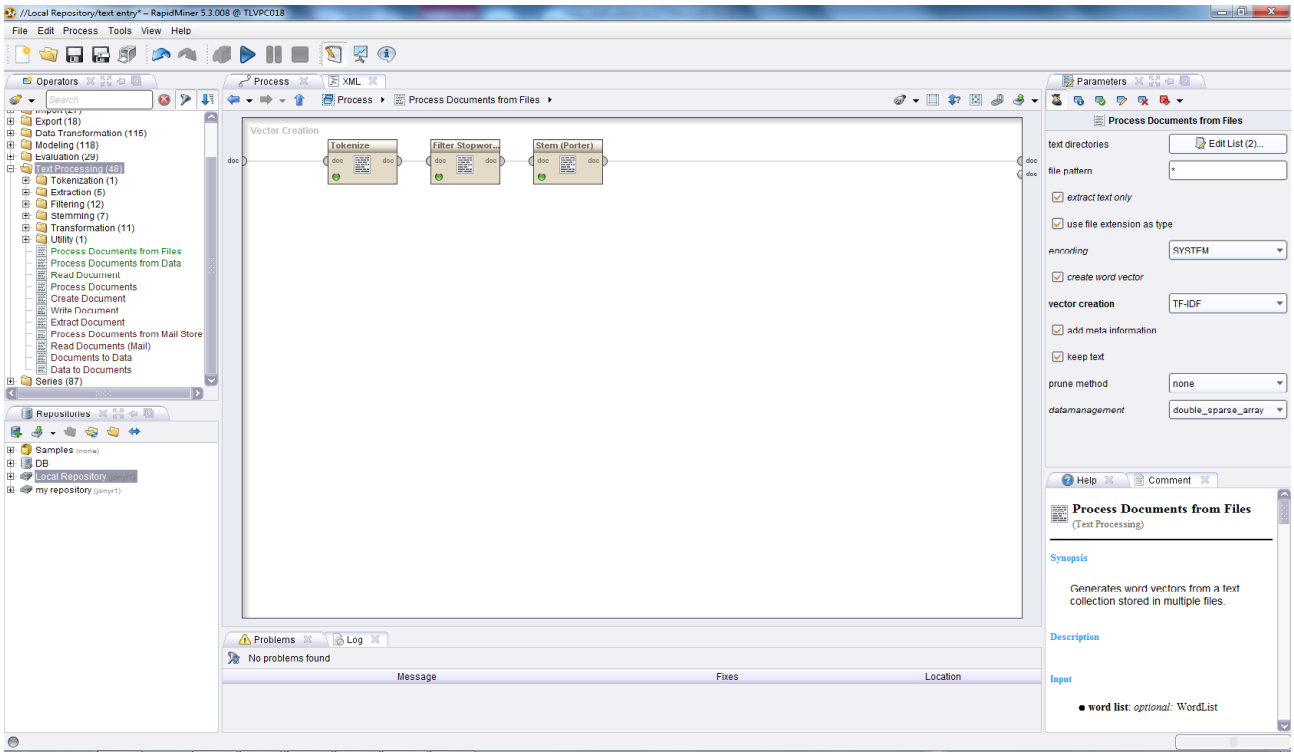

Figure 30: Sub-process of preprocessing task 4 on RapidMiner

Appendix 6B: KNIME

Task 1

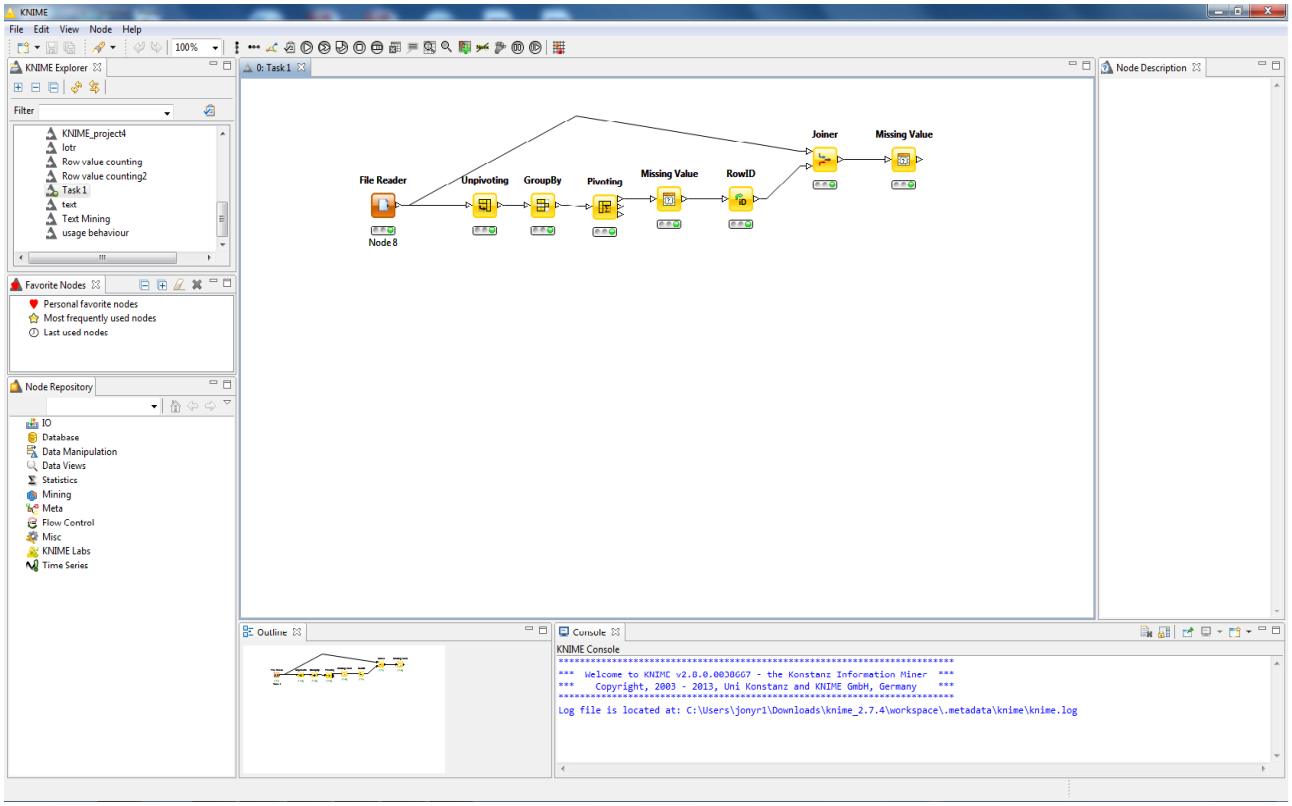

Figure 31: Preprocessing task 1 on KNIME

Task 2

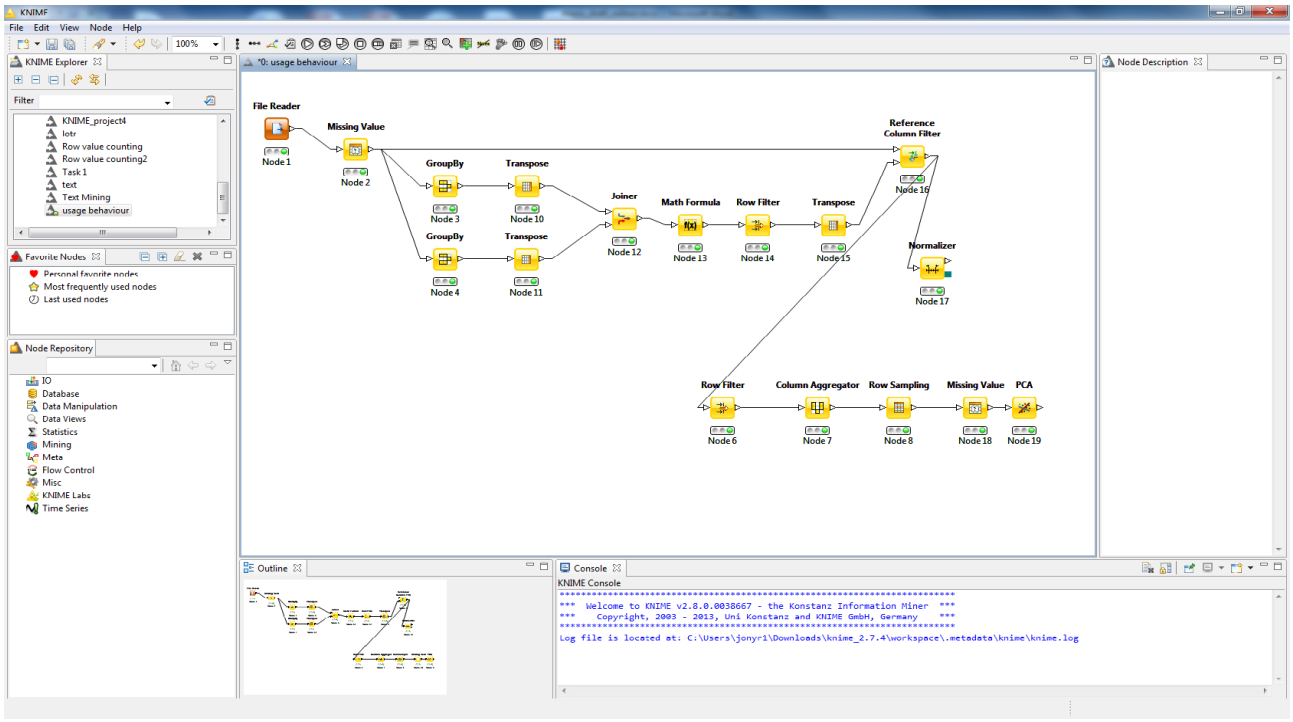

Figure 32: Preprocessing task 2 on KNIME

Task 3

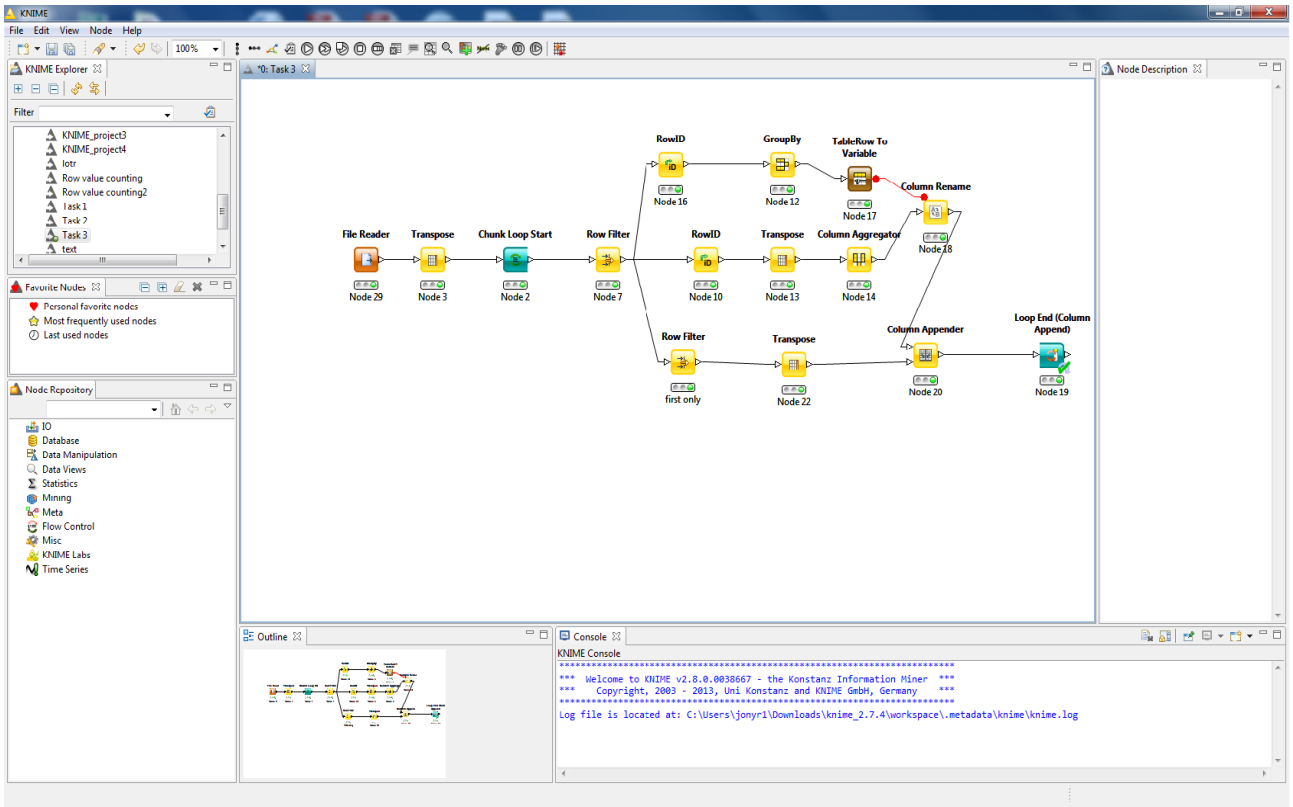

Figure 33: Preprocessing task 3 on KNIME

Task 4

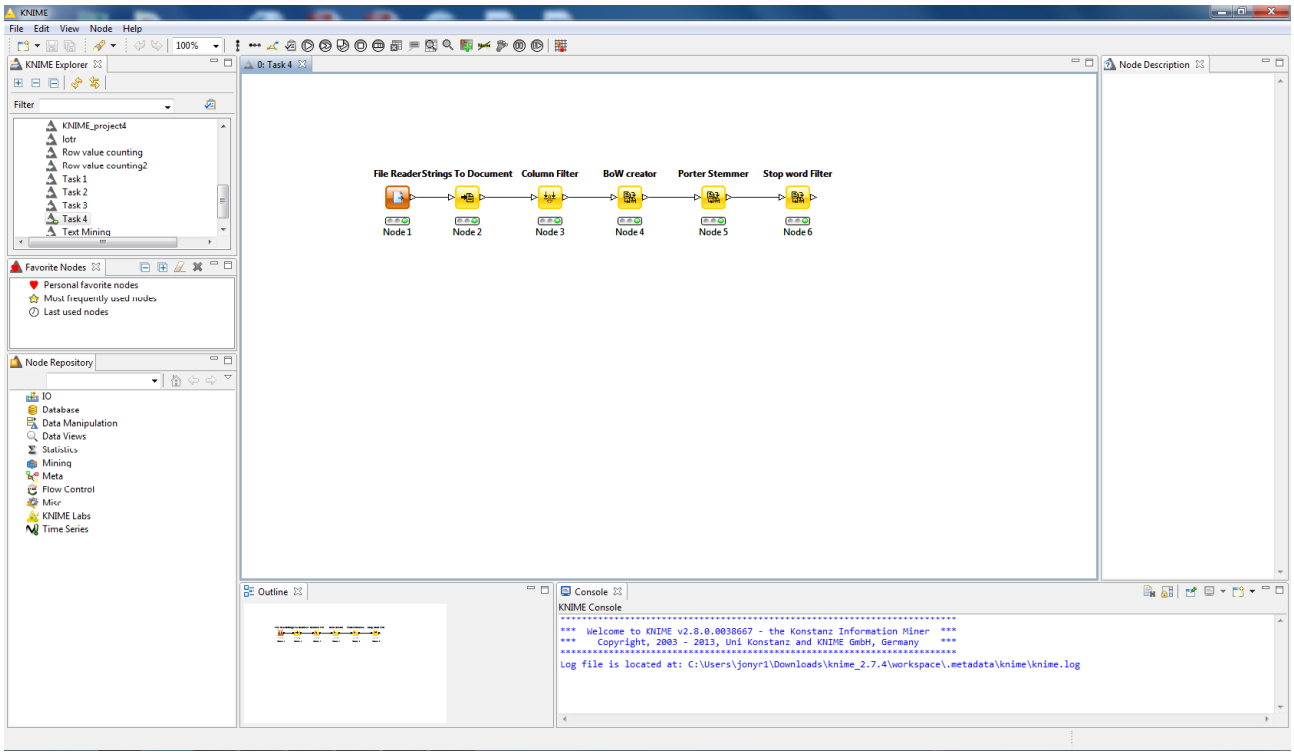

Figure 34: Preprocessing task 4 on KNIME

Appendix 6C: Orange

Task 1

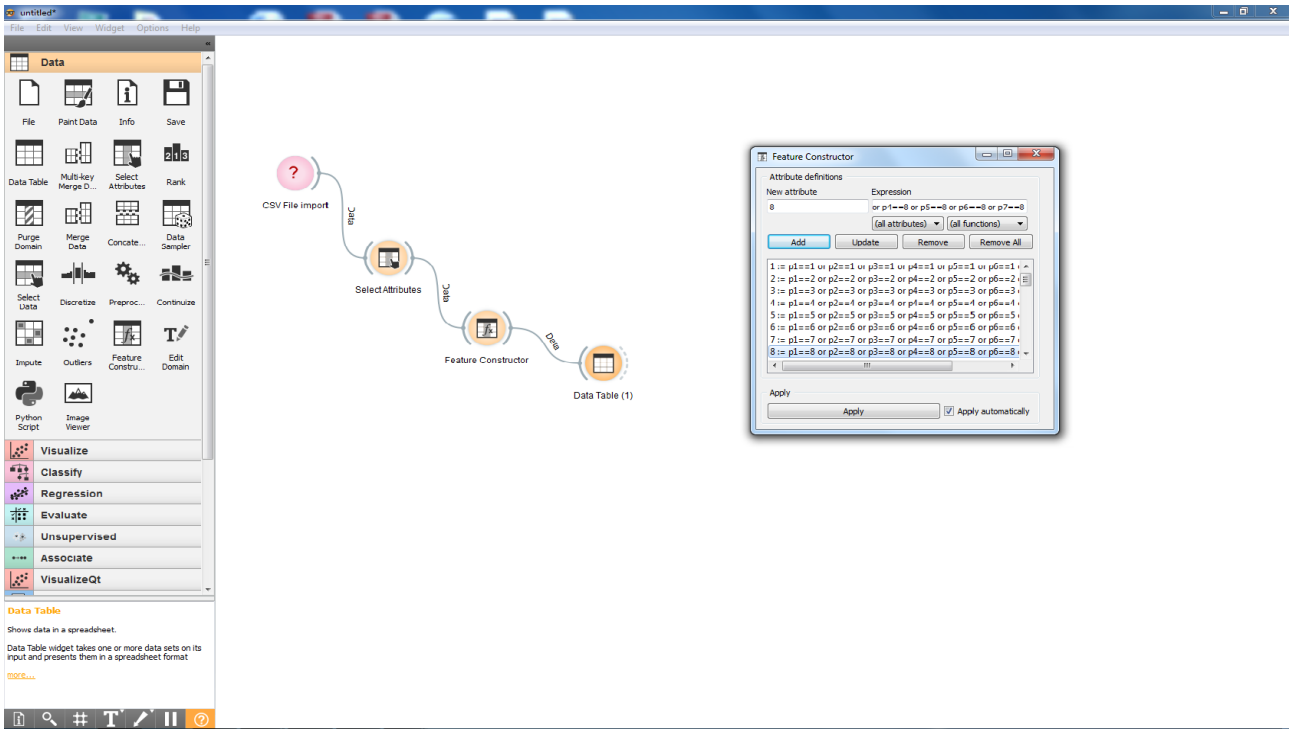

Figure 35: Preprocessing task 1 on Orange

Task 2

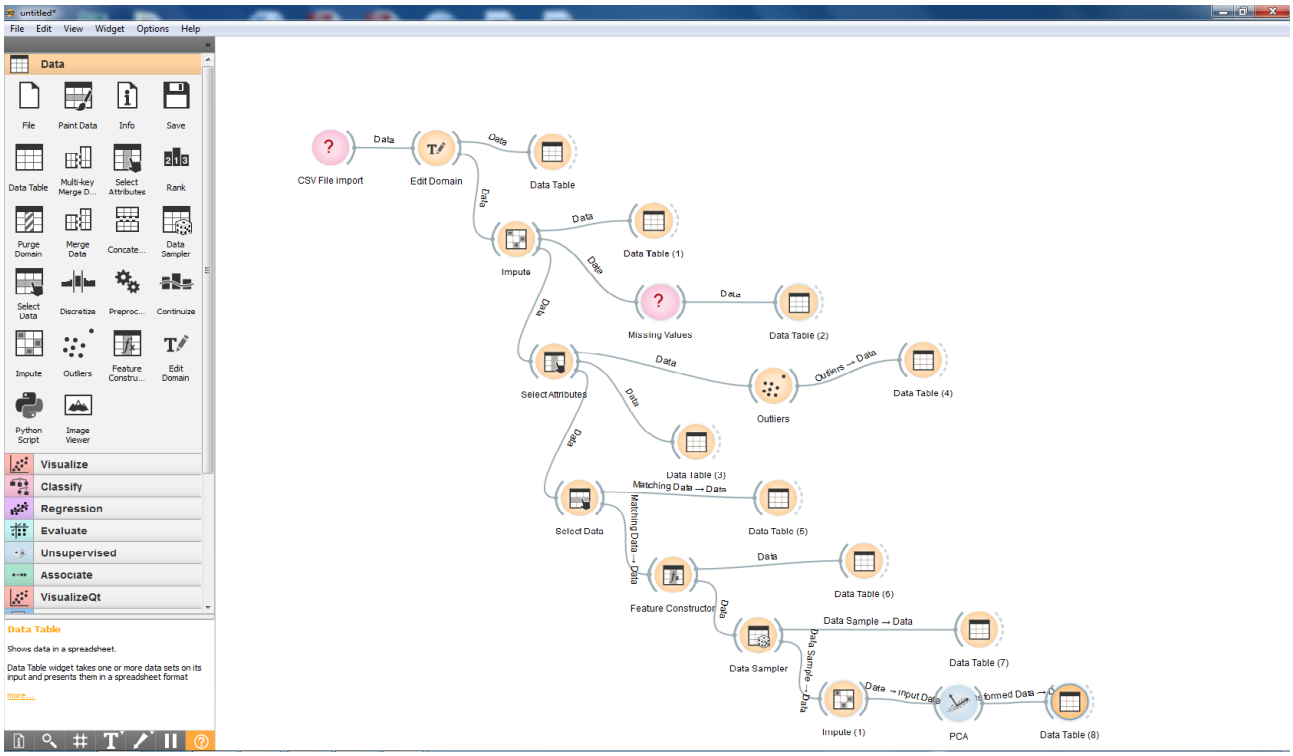

Figure 36: Preprocessing task 2 on Orange

Task 4

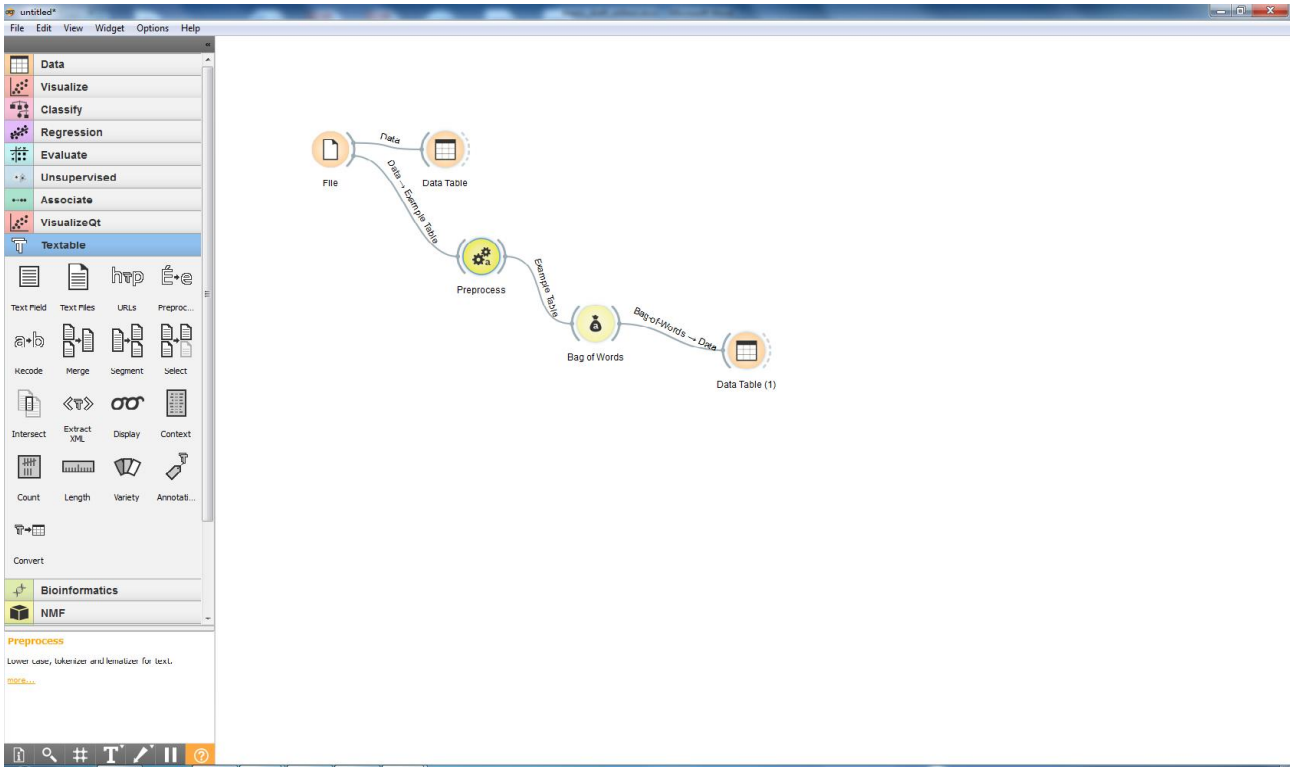

Figure 37: Preprocessing task 4 on Orange

Appendix 6D: IBM SPSS Statistics

Task 1

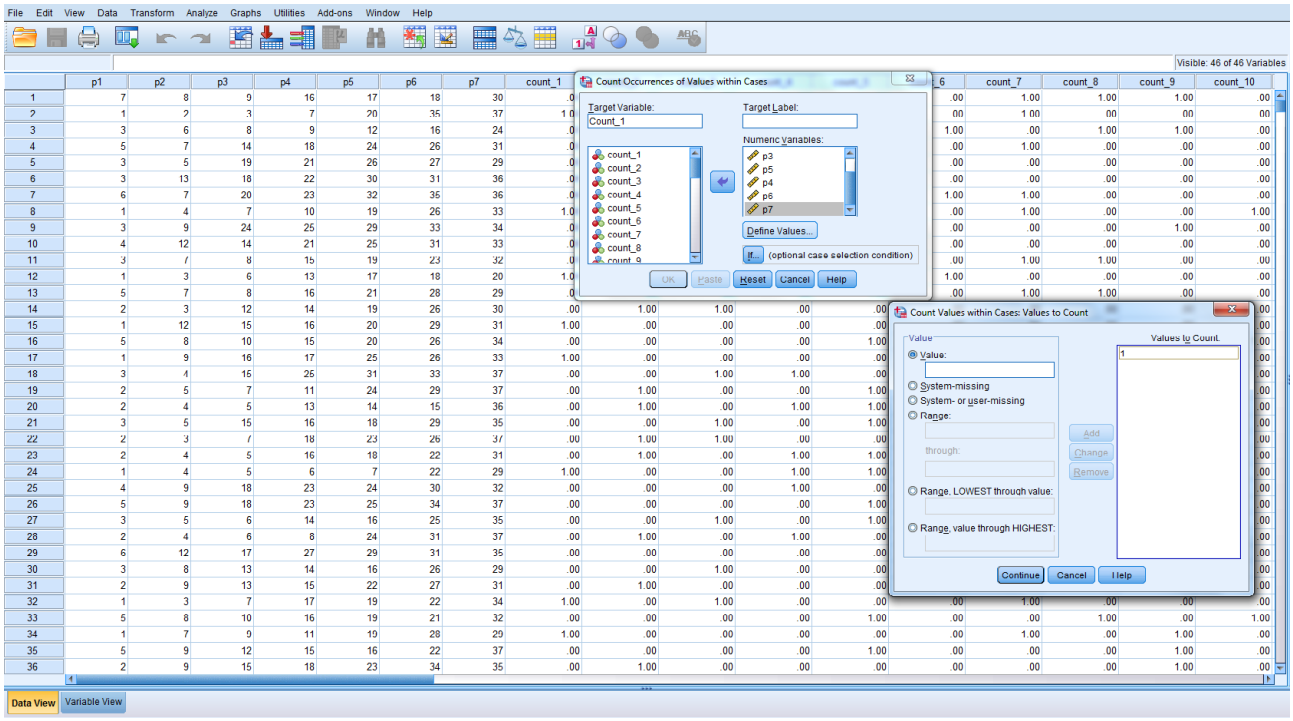

Figure 38: Preprocessing task 1 on IBM SPSS Statistics

Task 2

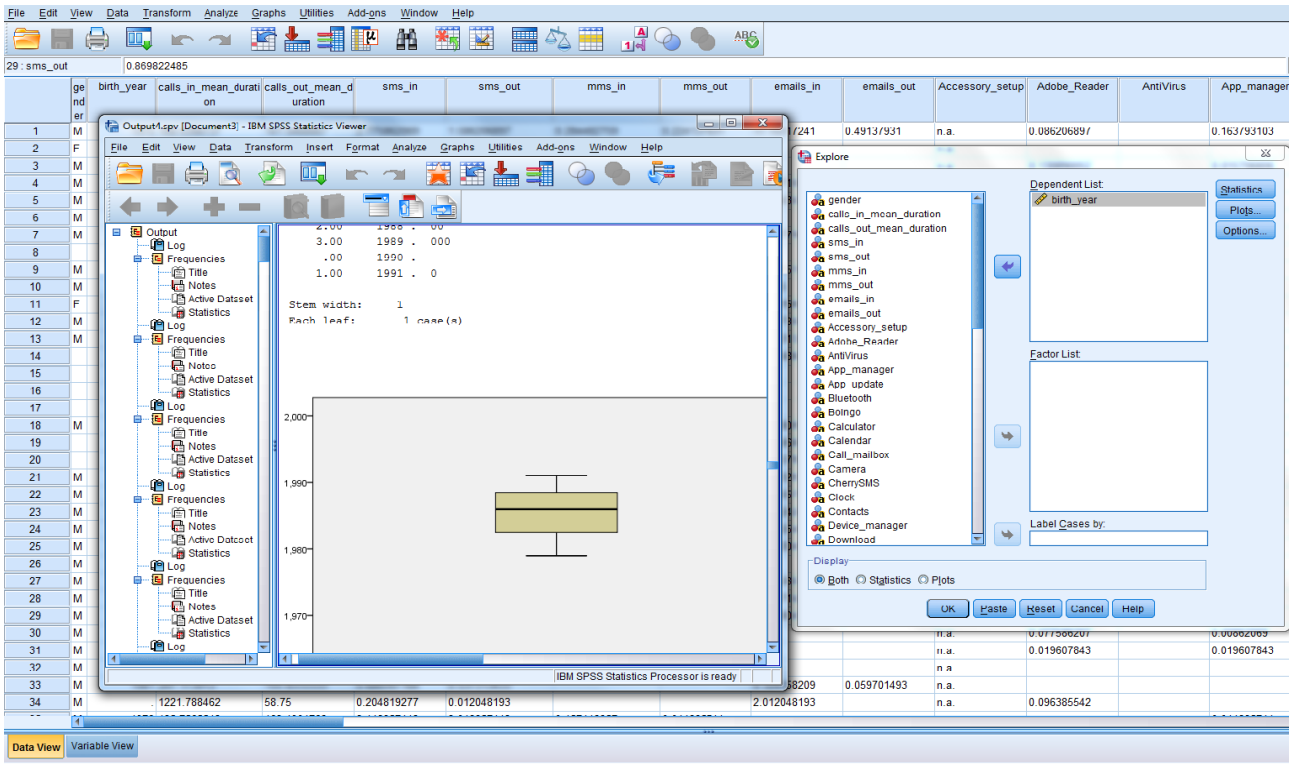

Figure 39: Preprocessing task 2 on IBM SPSS Statistics